THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF INFORMATION SCIENCES & TECHNOLOGY

USING STUDENT CHARACTERISTICS FOR MACHINE LEARNING MODELING IN HIGHER EDUCATION ALUMNI GIVING

DOMINIC MIRABILE
SPRING 2015

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree Electrical Engineering
with honors in Information Sciences & Technology

Reviewed and approved* by the following:

Vasant Honavar
Professor and Edward Frymoyer Chair of Information Sciences and Technology
Thesis Supervisor

Andrea Tapia
Associate Professor of Information Sciences and Technology
Honors Adviser

* Signatures are on file in the Schreyer Honors College.

ABSTRACT

In understanding complex relationships between variables and outcomes, the field of machine learning can be helpful for predicting and classifying new instances based on past data. This study applies a variety of machine learning algorithms to higher education alumni giving and compares the performance of each method in terms of classification accuracy and practicality for use. With increased reliance on private philanthropy, higher education must optimize their development efforts, which begins with the ability to identify likely donors. The models leverage the predictive power of student data to examine the link between student experience and alumni giving and to ensure the model could be applied to future classes using the standard set of collected data. The supervised machine learning methods employed in this study are: Naives Bayes, Decision Trees, Random Forest, Boosting, Bagging, and SMO Support Vector Machines. They represent diverse, widely used methods for classification and provide insight into the broader behavior of alumni giving.

The results of the classifiers were largely consistent in terms of classification accuracy and area under ROC curve—suggesting that the modeling had reached the highest accuracy possible for this specific dataset. Still the machine learning implementations perform significantly better than chance and offer a valuable tool in focusing engagement efforts on likely donors. This study contributes to the related literature by offering a practical application of several types of machine learning in a growing field where data is available and rich. It supports the importance of collecting data about potential alumni donors while they are students and encourages the further integration of machine learning into prospecting techniques.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge several key individuals that have encouraged my critical thinking, supplied me with the tools to succeed, and challenged me on this journey. To Dr. Vasant Honavar, my supervisor and esteemed professor, I can only hope that one day I have the experience and credibility that you possess. Your expertise and willingness to help were invaluable to me this past year. To Phil Manning, Associate Director for Analytics, your role as a coach and analytics professional not only provided me with the data to make this possible, but real-world insights to ground my research in your best practices. To Rod Kirsch, your enthusiasm to connect me with DDAR professionals and encourage my intellectual curiosity gave me the motivation to pursue and complete this study. I hope it will contribute in some way to the great work done by so many in the division for the betterment of Penn State. To Dr. Andrea Tapia, my honors advisor, I have learned so much about myself and the world we live in because our paths have crossed. I'll always cherish the passion you bring to the classroom and your students. To Cheryl Kaplan, my mentor and friend, your support of me in all my endeavors has changed me in profound ways during our short time working together. The list of lessons I take with me have certainly influenced my academic work, but more importantly, my character and charted path in life.

Finally, to all of my friends and family who have shown me how high I can reach but kept me square on the ground, no words will ever be enough to truly say thank you.

Introduction

Machine learning provides a method for visualizing and classifying large, complex datasets for a wide range of problems. While traditional statistical methods can generate the probability of an attribute predicting a certain outcome, machine learning algorithms generate classifiers that can computationally tease out the relationships between variables and classify new instances with their sets of attributes. This study leverages the theory and methodologies of a broader shift in the field of data mining from purely statistical analysis towards machine learning. The rapid adoption of machine learning methods is driven by both the increase in digital data collection and the development of robust algorithms with performance demands that have only recently become realizable. This study employs several well-known classifiers using the data mining tool, Weka, to compare performance across numerous methods.

The subject area of this thesis, alumni giving in higher education, is an appropriate problem domain for a number of reasons. Large colleges have hundreds of thousands of college graduates, yet only a small fraction of the population donates to their alma mater. Therefore, the ability to identify alumni that are likely to give is a key organizational imperative. Further, the reasons for giving are complicated and varied—characterized by an often unstated combination of factors that ultimately drive an alumnus to donate. While many studies have attempted to demonstrate correlative relationships between specific features and alumni giving, this type of modeling is often incomplete because of the complexity of decision making in philanthropic giving. Building a classifier for this problem set would greatly improve the identification of alumni by university development organizations, a process which is often ad hoc and intuition-based.

The study is conducted in consultation with the Research and Analytics unit for the Division of Development and Alumni Relations at Penn State University. Penn State is a large, public, research institution with several similar peer institutions nationally. The challenges of alumni prospecting and data mining are not unique to Penn State and the study is designed to be as generalizable as possible for similar peer institutions and nonprofits.

A defining element of this study is the choice of features used for modeling. While development professionals and researchers tend to focus on wealth estimates and survey data for higher education prospecting, this study restricts the attribute list to almost entirely characteristics collected while the student attended the university. This approach allows for the partial examination of student experience and how it relates to alumni giving, but it also builds a model that is applicable to all graduates of the college. In other words, the model is not contingent on the purchasing or otherwise collecting additional data from alumni post-graduation.

The research questions addressed in this study are as follows:

1) Do student characteristics alone provide enough predictive power to generate a machine learning classifier for alumni giving that performs better than chance?
2) Which classifiers have the highest performance set on this dataset and might be generalizable to this problem space?
3) How could implementing this type of model improve development efforts through targeting alumni likely to give?

To answer these research questions and build a reasonable solution, a review of the problem space and the data available will inform the selection and trial of several machine

learning classification algorithms. All results will be compared not only for classification accuracy but interpretability of results for use by development professionals.

Chapter 1

Literature Review

Higher Education Philanthropy

Higher education institutions have become increasingly reliant on non-tuition sources of income to finance annual operating budgets and capital expenditures for improvement (Monks, 2002). With the widespread decrease in government funding to higher education and simultaneous increase in costs of operating, private philanthropy has become an area of both opportunity and necessary investment for universities. While the size and performance of these development offices may vary, the primary target is alumni who have a strong connection through their college experience and a desire to strengthen or give back to the institution. In almost every case, the large number of alumni prohibits the engagement of all potential donors in effective solicitation, either through mailing campaigns or personal interactions by development officers (Lindahl & Winship, Predictive Models for Annual Fundraising and Major Gift Fundraising, 1992). To increase performance, and therefore revenue, development professionals must optimize the efficiency of their efforts through focused appeals.

This literature review will summarize and examine three relevant aspects of alumni giving: a framework for motivations to give, methods for prospecting and identifying alumni donors, and the use of machine learning in this type of predictive modeling.

Theoretical Framework for Higher Education Giving

To link individual characteristics to alumni giving, it helpful to understand the spectrum of reasons for giving from both a psychological and sociological perspective. These reasons can range from social benefits to satisfaction with the use of funds. In the case of academic philanthropy, one of the most common frameworks is a basic utility maximization framework, which suggests that each donor derives

some utility or satisfaction from giving to his or her alma mater (Weerts & Ronca, 2009). The model also suggests that increased giving stems from increased utility.

These motives are surveyed and summarized by Bekkers and Wiepking in their literature review on generosity and philanthropy (Bekkers & Wiepking, 2007). Their research looks at empirical trends in order to answer the question "Who gives how much?" and then leverages extensive surveying to understand why people give. The eight most important forces that drive giving and are widely accepted across philanthropy are: (1) awareness of need; (2) solicitation; (3) costs and benefits; (4) altruism and impure altruism; (5) reputation; (6) psychological benefits; (7) values; (8) demonstrated efficacy (Bekkers & Wiepking, 2007). For the purposes of this study, the mechanisms that may be at play in giving as predicted by student characteristics are: awareness of need, demonstrated efficacy, and costs and benefits. Each of these three motives relate to the donor's prior experience as a student and may be therefore represented directly or indirectly in student data.

Many studies on charitable giving have shown that awareness of need is a critical precursor to giving (Berkowitz 1968; Berkowitz and Daniels 1964; Schwartz, 1975). Additionally, donors must also feel that their giving makes a difference to the organization. During their time in higher education, students are not only made aware of the need and the institution's efficacy, they directly experience it. With the high fraction of all students at large public institutions being supported by scholarship aid or engaged with student programs and organizations funded through philanthropy, this mechanism is worth exploring.

Costs and Benefits as a giving mechanism has several interpretations when applied to higher education philanthropy. While the costs of giving are a limiting factor for donors generally, the benefits of giving may be more relevant to predicting giving based on student experience characteristics. Similar to grateful patient philanthropy in hospitals, alumni-giving is frequently viewed as a desire to repay the university for the education or benefits received (Leslie and Ramey 1988). These "benefits" could be manifested in a number of ways. Frequent contact with faculty (Monks, 2003), strong academic

reputation (Cunningham and Cochi-Ficano, 2001), and mentoring programs (Clotfelter, 2003) have all been positively correlated with alumni giving. This mechanism therefore suggests that donors would give at levels corresponding to the perceived quality of their college experiences and value. Several studies have also gone further to suggest that undergraduate extracurricular involvement and engagement may be a factor in alumni giving (Miller & Casebeer, 1991).

Still others studies cite obligation to society (O'Connor, 1961), attitudes about the institution (Korvas, 1984) (Hall, 1967), satisfaction with collegiate personal and social growth (Thomas & Smart, 2005), or participation in alumni activities (Caruthers, 1971) as dominant factors (Miller & Casebeer, 1991). However, in almost every case, there are mixed results. The splintered nature of research on this topic shows no consensus on the best way to classify alumni as likely donors and underscores the complexity of alumni giving as well as the need for robust analysis.

Of course, this framework does not fully address the ability to give and is largely confined to motivations for giving. Certainly, motivations alone are not sufficient in understanding gift behavior because the capacity to give may very well be a limiting factor. However, beginning with specific motives for giving is important for deciding which donor characteristics to include in prospect modeling and will determine the limits of the model's predictive power. For this reason, the utility maximization and donor motive framework has been used as a starting point for formulating strategies for modeling, rather than the capacity to give.

Methodologies for Prospecting in Alumni Giving:

Because university development offices have limited resources and relatively large numbers of potential alumni donors, identifying alumni with a high likelihood of giving is the primary objective. Additionally, for many universities, the giving profile follows the conventional 80/20 rule, where 20% of

donors give 80% of the total gift dollars after being cultivated from an early age (Carolina Angel, 2015). Therefore the secondary objectives are identifying donors likely to contribute a significant or major gift and building a pipeline of young alumni donors. A well-established area of interest and effort in both nonprofit and commercial sectors is the idea of prospecting, or searching in a systematic way for individuals with a high probability to respond favorably. For commercial sales, professionals may create algorithms or reports that identify potential customers based on their web or surveyed behavior. For nonprofit fundraising, major gift fundraisers may use wealth estimates and past giving as a means to prospect potential large donors. For many higher education institutions, the selection process for determining the best individual potential donors is ad hoc and based on intuition rather than predictive modeling. (Lindahl & Winship, 1992). Basic approaches involve selecting one characteristic, i.e. estimated wealth based on census data, and confining the solicitation pool to individuals within a certain range. The limits of this approach become obvious when considering all of the additional factors that could influence the decision to make a gift besides wealth. This example represents two flaws with rule-based prospecting: eliminating potentially useful information from the analysis and assigning arbitrary thresholds or ranges not necessarily backed by evidence.

When applying more advanced methods, universities can incorporate significantly more information into the modeling, generate predictive weights for each feature in a set of attributes, and tailor the method appropriately to the dataset. These statistical models have significantly advanced the field of higher education philanthropy prospecting, but still represent an incomplete approach to donor prediction as discussed below.

Statistical Techniques: Regression Modeling

The common approach to predictive problems is statistical modeling. An extensive literature review of alumni giving for higher education yielded, almost exclusively, studies that used statistical

techniques to assign weights to features and calculate a probability that an individual with that characteristic would make a gift. This process is achieved through regression, which estimates the relationships between variables by calculating how outcomes vary when a dependent variable changes. Specific to alumni giving, the primary techniques included logit, probit, and tobit regression modeling.

Logit models, or logistic regression models, predict a dichotomous outcome by modeling the log odds of the event as a linear combination of the predictor attributes. With dependent variables that are "flags" (either a zero or one), logit regression is often used as a regression model because it has the benefit of using log, rather than linear, relationship modeling. Since many donor characteristics are represented with dummy variables (gender, scholarship recipient status, first generation college student) so they can be simplified in donor databases, logit models can assign a predictive probability factor to each one. For example, Gaier, 2005 uses a logit regression to determine the odds of alumni involvement based on satisfaction with undergraduate academic experience as characterized by 17 independent variables.

While Gaier was able to assign a weight for each factor in determining alumni giving, he suggested that "subsequent research should focus on the integration of these factors… [combining] the interaction of the academic system and social system (Gaier, 2005). This approach would more comprehensively address the complexity of the giving relationship and provide a mechanism to assess the complete donor—rather than just isolated characteristics of a donor. Lindalh and Winship also employ a logit model to characterize interactions of features in major gift donor prediction and finds that past giving is the best predictor for future giving (Lindahl & Winship, A Logit Model with Interactions for Predicting Major Gift Donors, 1994). Integrating additional data into a giving history study would only improve confidence values and better characterize alumni giving.

Probit models are almost identical to logit models in approach with the primary distinctions being the ease of interpretation and the parameter ranges. Tobit models offer the ability to set a minimum or maximum threshold for numeric attributes, which can be useful when examining stratified alumni giving.

These examples, among many others, demonstrate that regardless of the slight variations between regression methodologies, models that only generate weights or probabilities for individual attributes may not be the most comprehensive and actionable approach to prospecting in higher education philanthropy.

Machine Learning Methods for Higher Education Prospecting:

While statistical models comprised the majority of methods used in alumni donor prospecting, some studies have turned to machine learning classification to advance the understanding of factors at work in higher education philanthropy. The Classification and Regression Tree (CART) methodology was employed by Weerts & Ronca (2009) and suggested for further research by Key (2000). Classification trees can be helpful in producing an accurate classifier for use on new observations as well as uncovering the predictive framework of the problem (Russell & Norvig, 2009). While decision tree models like CART will be discussed in Chapter 3: Methodologies, the overall concept involves beginning with a root node attribute, splitting along the options for that classifier, and repeating that process until reaching an outcome where all observations that arrive at the "leaf" exhibit the path's characteristics. The stated reasons for using CART in addition to preliminary statistical analysis were: (1) the ability to handle collinearly dependent attributes and missing data, (2) incorporation of a cost function for incorrect classification, and (3) the construction of a robust algorithm to classify new donors with a validated correct classification percentage. In addition to providing an actionable framework for classifying new potential donors, their specific dataset revealed that key characteristics in predicting alumni donors relate to "attitudes, beliefs, income, keeping in touch with [the university], and the number of degrees from other universities" (Weerts & Ronca, 2009).

Another study on enhancing fundraising success with custom data modeling began with a probit analysis and achieved mixed results. In analyzing the dataset and possible future methods, he suggests

using Chi-Squared Automatic Interaction Detector (CHAID) or CART in future research. CHAID is another decision tree machine learning technique based on adjusted significance testing of purely categorical variables. The tree is built by splitting on each variable as selected by the chi-square significance test of the variable's contingency table.

The demonstrated utility, success, and interest in applying machine learning algorithms to predicting alumni giving warrants the further exploration in this area. The key areas of interest where value could be added relate to: (1) the range of alumni attributes, (2) types of machine learning classifiers, and (3) supplementing current modeling techniques.

Features with High Predictive Power:

Whether using statistical methods or machine learning algorithms, the features – or data included on each individual—are equally important in determining the results and utility of the model. Prospect modeling typically leverages both internal data, such as past giving, gift potential ratings, and volunteer activity, as well as external data, including survey responses, wealth estimates, geodemographic coding, and employment history (Lindahl & Winship, 1994, "Logit Model with interactions for predicting major gift donors"). Models that leverage multiple donor features are superior for predicting future giving, but certain features often account for much of the predictive power. One example of a feature with high predictive power is past giving. This finding is reasonable, as past giving is a proxy for a whole host of other features related to giving, but it is not particularly valuable in identifying donors at the early stages of the giving lifecycle with little giving history. For these recent graduates, more data is needed for reliable classification.

In an extensive literature review, Taylor and Martin (1995) compile more than 30 doctoral dissertations that examine alumni giving and predictive features. The attributes were explored and cross-

referenced, and many suggested some level of mixed results. Combined with results of other studies, the attributes, correlation description, and cited authors most relevant to this study are listed below:

Table 1: Attributes Examined for Correlation with Alumni Giving

Attribute	Likelihood of giving	Cited Authors
Age	More likely	Haddad, 1986; Oglesby, 1991; Beerler, 1982; Bruggink & Siddiqui, 1995
Emotional attachment to alma mater	More likely	Leslie et al., 1983; House, 1987, Taylor & Martin, 1995; Diamond & Kashyap, 1997
Perception that university does not need gifts as much as other organizations	Less likely	Pearson, 1999
Satisfaction with educational experience	More likely	Oglesby, 1991; Shadoian, 1989
Satisfaction with undergraduate experience	More likely (Van Horn) Less likely (Miracle; Houes)	Van Horn, 2002; Miracle 1977; House 1987
Involvement in extracurricular activities	More likely (Shadoian; Oglesby; Miracle; Morris; Gardner) Less likely (Beeler; Miller & Casebeer) No predictive power (Grill; Kraus; Young & Fischer)	Shadoian, 1989; Oglesby, 1991; Miracle, 1977; Morris, 1970; Gardner, 1975 Beeler, 1982; Miller & Casebeer, 1990 Grill, 1988; Kraus, 1991; Young & Fischer, 1996
Participation in fraternity or sorority	More likely (Bruggink & Siddiqui) Less likely (Okunade, Wunnava, & Walsh)	Bruggink & Siddiqui, 1995; Okunade, Wunnava, & Walsh, 1994
Participation in special interest group	More likely	Taylor & Martin, 1995
High Income	More likely	Weerts & Ronca, 2009; Leslie & Ramey, 1988; Wastyn, 2009
Marital Status	More likely	Sun et al., 2007, Monks, 2003
Relatives/children attended alma mater	More likely	Weerts & Ronca, 20209; Wunnava & Lauze, 2001
Receiving need-based grants	More likely	Hoyt, 2004
Receiving student loans	Less likely	Monks, 2003

Despite providing whether an attribute is positively or negatively correlated with giving, this compilation of studies is not very helpful to development professionals. If a fundraiser were to filter all alumni with one of these characteristics designated with a "more likely to give" label, the subset would still be too large for personalized solicitation. Additionally, an alumnus can be identified with a number of these features, likely with conflicting predictive outcomes. More robust techniques for filtering and prediction are needed to maximize the efforts of development professionals, but the number of alumni, diversity of experience and affinity, and competing attributes make this problem both challenging and worthy of exploration.

A related challenge for development professionals is collecting data on potential donors. Many resort to purchasing wealth estimation data or surveying alumni to build out their database. Because many of the features commonly used in modeling are related to post-graduation characteristics, there are few studies of before-graduation-characteristics and even fewer that leverage large and reliable datasets. One such study separated variables for predictive modeling into "before graduation" and "after graduation" categories to assess whether a donor would fit into an annual gift or major gift designation. The model included a two-stage Heckman-style regression on a donor utility equation of the probit results (Lara & Johnson, 2008). The results of annual gift and major gift estimation results for before graduation features were not as definitive as post-graduation characteristics and were fundamentally limited by the student data available. For the annual gift model, the number of collegiate honors received, class officer designation, and greek (fraternity or sorority) affiliation predicted both a likelihood of annual giving and an increased giving level. Number of varsity sports played, number of intramural sports played, number of student activities, and playing on the varsity men's hockey team predicted a lower likelihood of giving as well as a lower average giving amount. The major gift model linked varsity sport participation and high student activity participation to being less likely to give at the major gift level, but failed to produce statistically significant results for many of the other features due to the low number of major gift donors available (only 55 with adequate information). Additionally, a confounding factor of the major gift

analysis may be that major gift donor populations are statistically much older than annual gift donors and there is significantly less data available on older donor populations who predated electronic capturing of student data. Another study on young alumni giving patterns across 28 universities concludes that "even conditional on income, advanced degree attainment, and overall satisfaction with one's undergraduate experience, the major field of study are significant determinants of alumni giving" (Monks, 2002). By running a tobit regression on observations sorted into major categories, the experiment suggested that graduates in the fine arts or nursing gave significantly less and graduates in history gave more than their counterparts in the humanities.

While both of these studies provide validated examples of before-graduation attributes predicting alumni giving, contradicting or mixed results suggest that it is very difficult to generalize the predictive power of features relating to a student's experience. This implication seems intuitive as student experience varies widely across students, across universities, and across time. While this claim may seem discouraging for development professionals, it actually supports the use of machine learning methodologies for this particular classification problem. The size of alumni databases and largely nominal feature sets allow the construction of machine learning frameworks built on past data that can characterize new individuals as likely donors or non-donors.

Machine Learning: Supervised Classification Problems

Where most studies in the area of higher education donor prospecting employ statistical methods or regression techniques based in probability theory, this study applies machine learning methodologies to this complex problem. The differences, both in approach and results, is worth exploring in the context of the literature to underscore the novelty of this approach. In many of the studies reviewed, survey or demographic data was collected and analyzed using logistic regressions or other probability based techniques. The results offered a coefficient or "odds" that an alumnus with that characteristic (i.e. high

satisfaction with undergraduate experience) would make a gift to the university. Armed with this information, development officers could feasibly filter the alumni population by individual features that had a high probability of giving and solicit those groups. Statistical methods alone do not have the power to take a new observation (i.e. an alumnus with a set of features) and assess that alumnus' likelihood of giving or categorize him into a donor or nondonor category. That is to say, it is very difficult to assess the combination of features and their impact on giving likelihood. Machine Learning algorithms aim to provide this type of robust solution.

The theoretical basis of machine learning is the ability to generalize from experience; in this case "experience" is a dataset with features and known outcomes. The automated computing procedures that process this data are rooted in logical or binary operations that allow the program to learn a task from the series of rich examples. The field is a subset of a larger subject area known as artificial intelligence, and has applications reaching from facial detection and spam filtering to medical diagnosis and weather prediction. Complex problems with high feature counts benefit from machine learning because the utility of the model grows by increasing the amount and richness of data. The goal and driving factor for developing machine learning algorithms is to predict an outcome with equal or better credibility than a human counterpart, and at a larger scale than would be feasible for humans to reasonably assess (Witten, Frank, & Hall, 2011).

The subset of machine learning relevant to this study is supervised machine learning, where the instances are labeled and categorized in some way prior to analysis. The data in development databases must be categorized by design and the outcomes are also labeled (i.e. donor, nondonor). Additionally, the primary family of machine learning techniques employed for this study is classification algorithms; which are defined by "the construction of a procedure that will be applied to a continuing sequence of cases, in which each new case must be assigned to one of a set of predefined classes on the basis of observed attributes or features." (Michie, Spiegelhalter, & Taylor, 1994).

The elements needed for a classification problems, according to a compilation of various classification techniques by Michie, Spiegelhalter, & Taylor (1994), are listed below:

1. The relative frequency for which the classes occur in the population of interest, expressed formally as the prior probability distribution.
2. An implicit or explicit criterion for separating the classes: an underlying input/output relation that uses the observed attributes to distinguish a random individual from each class.
3. The cost associated with making a wrong classification.

This structural design is almost exclusively informed by the background information of the problem and the characteristics of the data. The resulting learning algorithm is then applied to a subset of the dataset—known as the training dataset—to generate the classification "rules" or framework. Once the designer is satisfied with the accuracy of the classification rules in fitting the training dataset, the classification framework is applied to the testing data as shown in the figure below.

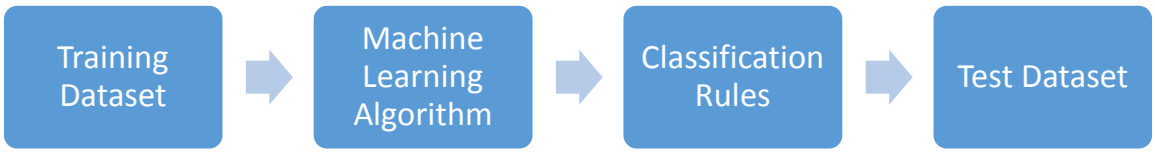

Figure 1: Classification process

There are common challenges with supervised learning algorithms that can affect the fitting or accuracy of the model. Many of these issues are encountered in the pre-processing phase and can include: incorrect or missing values in the dataset or irrelevant input features skewing the analysis. At the onset, it is important to determine the scope of the missing or incorrect data and how it might affect the resulting

model. Another challenge with supervised machine learning lies in the selection of the specific learning algorithm. Examples include logic-based techniques, perception-based techniques, and Bayesian models; where each family of classification methods has particular strengths and limits. Determining the type of model to use often depends on the dataset as well as the requirements by the user. It may be impractical to use a certain model if the data is not rich enough or large enough. Similarly, depending on the application, the user may value certain metrics for the algorithm differently. The four metrics that distinguish models are: accuracy, speed, comprehensibility, and time to learn (Russell & Norvig, 2009). For example, in machine learning problems relating to financial transactions, accuracy might be valued much more than comprehensibility to a human user, whereas in a medical diagnosis, the opposite might be true. On one hand, understanding these considerations and metrics in scoping the problem is essential to determining the appropriate model. On the other hand, it may be difficult to understand how each method will score according to these metrics. In that case, it is often recommended to simply try each one separately and compile lessons learned.

In conducting this literature review, it became clear that the use of classicization machine learning algorithms for donor prediction is either not done in industry or not well published in research communities. Additionally, the increasing use of machine learning across a wide array of application areas provides an extensive repository of studies for how these types of methods have been applied to similar problem domains. A dual understanding the theory, process, and landscape of machine learning classifiers and higher education alumni giving is necessary to complete this study. The types of models employed and the novel use of student attributes adds two elements to this subject area that will extend research in this area and supplement the working model for development professionals.

Chapter 2

Data

The dataset for this study is taken directly from the Division of Development and Alumni Relations (DDAR) database at a large, public research university with permission for use in this experiment. The preprocessing steps are described below along with a characterization of the scope and limits of the data. The attributes included in the machine learning modeling in Weka is only a subset of the larger set of features collected for exploratory analysis. Details about frequency of giving and total lifetime giving were analyzed, but not included in modeling. The purpose was to restrict the modeling attribute list to features collected about students during their time of enrollment to determine the predictive power of such a classifier.

Included Attributes

The full feature list retrieved from the database is listed below. Decisions on which attributes to include in modeling were made based on (1) the desire to restrict the modeling to student characteristics and (2) the evaluated integrity of the attribute in terms of accuracy and richness.

Table 2: Full Attribute List

Attribute	Type of Attribute	Additional Comments
AlumID	Database	
EnrollmentState	Demographic	Home state at time of enrollment
Age	Demographic	
BirthYR	Demographic	Redundant verifier for Age
GENDER	Demographic	
EMAIL_IND	Database	Indicating if email address is known. If Y, likely used for contact
DegreeYear	Demographic	
MAJOR_CODE	Database	
MAJOR	Academic	
PREF_CLASS_YEAR	Demographic	Self-reported, redundant verifier for DegreeYear
PREF_SCHOOL_CODE	Database	
PREF_DEGREE_LEVEL	Academic	Self-reported designation for PSU degree level
HOME_CITY	Demographic	
HOME_STATE	Demographic	
HOME_ZIPDIST	Demographic	Used to determine local college students (less than 15 mi from campus by zip code)
PSUDonor	Class	Outcome class
AttendCampus	Academic, Student Experience	Denotes attending a commonwealth campus prior to or instead of University Park (Main) campus
CampusGraduate	Academic, Student Experience	Qualifies students attending commonwealth campuses: Y - Never attended University Park (Main) campus, N - Attended University Park for portion of degree
FirstDegreePSU	Academic	
MultiPSUDegrees	Academic	Verifier for PSU degree level
ActiveAlumMember	Alumni Activity	Denotes if membership in the Penn State Alumni Association is currently active
YrsDblFirstGf	Giving	Years until lifetime giving doubles first gift amount
LifetActualAmt	Giving	Lifetime giving amount
HHCreditAmt	Giving	Household giving amount (includes spouse giving)
FirstTHON	Giving	Denotes if first gift was to THON, a cause based student philanthropic effort
FirstAth	Giving	Denotes if first gift was to Athletics, an independent and reward-based giving mechanism
PercentGiveATH	Giving	Calculates percentage of total giving amounts designated to Athletics
PercentGiveTHON	Giving	Calculates percentage of total giving amounts designated to THON
NumAlloc	Giving	Number of Allocation codes used by donor for lifetime giving
AvgGf	Giving	
NumsOfGive	Giving	
GiftFreq	Giving	Calculated number of gifts/year
YrsToFirstGF	Giving	Years between earliest graduation year and first gift
YrsOfGive	Giving	Number of years gift is reported

FirstGiftAmt	Giving	
LastGiftAmount	Giving	
FiveYrsSign	Giving	Last five years giving marker sequence i.e. 11111 denotes a gift each of the past five years
NumStudAct	Student Experience	Number of activities coded for participation
StudActTHON	Student Experience	Participation in THON, a large philanthropic activity
StudActGreek	Student Experience	Participation in fraternity of sorority
StudActLL	Student Experience	Employed in LionLine, a student calling program to solicit university gifts
Scholarship_Marker	Academic	Denoted recipient of a need-based or academic scholarship
OrgType	Student Experience	Specifies type of extracurricular involvement, given Stud_Act_Marker positive
Stud_Act_Marker	Student Experience	Denotes participation in a student activity
Athletic_Marker	Student Experience	Denotes collegiate athlete
VarsityAth_Marker	Student Experience	Denotes varsity athlete

The data incorporates every measured element of the student experience that is reasonably available to the study. Student Experience is characterized to varying degrees by: Stud_Act_Marker, Athletic_Marker & VarsityAth_Marker, AttendCampus, and Org Type. Demographic information and academic traits is contained in Age, GENDER, DegreeYear, MAJOR, PREF_SCHOOL_CODE, and HOME_ZIPDIST. Finally, the Scholarship_Marker designation identifies recipients whose student experience and degree was enabled by the behavior this study seeks to classify: philanthropic giving. While there are several other conceivable measures of the collegiate experience, the attributes used represent the extent of student-related information collected by the division. Given the mixed results of studies examining the link between student experience and alumni giving and the general lack of student information captured by development organizations, this study begins to uncover the value of student data in alumni gift prospecting. Additionally, while studies (Gaier, 2005) (Monks, 2002) (Thomas & Smart, 2005) labor over the cause of alumni giving, this experiment seeks to actually build a model for classifying new individuals based on student characteristics. Therefore the attributes used and the models built may not uncover the reason alumni give to their alma mater. However, the machine learning methods will demonstrate if those traits have any predictive power in classifying as donors and nondonors. Given that universities can easily

collect information on all individuals while they are students, these findings would further supplement and improve prospecting efforts.

Academic Major Data

Individual academic experience is characterized in this study by a student's academic college. Students from the same college may share similar career opportunities, interests, and experiences. A simple frequency count of donors by college for graduates from 2004-2009 reveals that certain colleges have significantly higher donor percentages as shown below. These discrepancies could also be explained, in part, by the relative size of the development units. The highest number of donors are held by Liberal Arts, Engineering, Health and Human Development, and Business, which are also in the top five development units in terms of size and giving goals.

Student Activities

Purely academic demographic data, such as major or college, is not rich enough to characterize a student's entire experience. Additionally there are studies that support out-of-class learning experiences and student activities as instrumental in social, personal, and professional growth. While the literature review compiles conflicting findings on the participation in student organizations in predicting alumni giving, few studies adequately explore the different types of experiences that are coupled with different types of student organizations. This study aims to uncover the predictive power in types of student organizations for alumni giving.

In determining the categories of student organization involvement, the university's internal designations were used as a starting place. Registered student organizations must specify their organization as one of 15 categories as part of a process enacted within the last ten years. The set of

student activities were sorted into these categories, but not all of the categories were used. Additionally, three notable categories were added: Mentoring Program, THON, and Penn State Pride Organization. Mentoring programs is a student activity reported through the academic colleges and while they are not organized through a student organization, they do impact the student experience. THON, although a philanthropic organization, was separated out because of its unique and somewhat dominating role in the university community. Not only is it one of the largest organizations, but it is a platform to generate gifts to the division, which are then allocated to a specific cause. Because of the unique philanthropic nature of the activity and its generation of donations for the division, participation is THON was decoupled from other designations. Finally, several organizations were placed into the category of Penn_State_Pride_Organization. While this may not be generalizable for all college campuses, several student organizations exist to perpetuate the history and traditions of the university and by doing so, create a deep connection between students and their alma mater. By isolating them within a separate organization type, the role of participating in a college-affinity based organization may be determined.

The number of students coded in development database for participation in student organizations is strikingly low. It is likely that full organizational rosters were not turned over to development for recording and that many general members were not coded. However, the relative frequency across the categories of student organizations seems reasonable. This consistency is likely due to the fact that the officers of each organization were recorded by default. While there are certainly past students that were involved, but not coded as such, it is also reasonable to assume that those coded were highly involved and had defining experiences in their respective organizations. The low number of recorded student activity is likely due to its reliance on either self-reporting or the exporting of student information from other systems housed in different divisions.

Table 3: Student Organization Types

Organization Type	Number of Instances
Academic_Club	150
Academic_Honors_Society/Program	644
Campus_Leadership	424

Global_Program	34
Greek	3340
Honors_Society	392
Mentoring_Program	461
Other	2274
Penn_State_Pride_Organization	897
Performing_Arts	294
Philanthropic_Organization	592
Professional_Development_Club	131
Publishing/Communications_Club	101
Religious	37
ROTC	129
THON	1904
Grand Total	**11804**

Athletic Involvement

The participation in athletics is included as an attribute for several reasons. The student experience of an athlete is unique, community-based, and demanding. It affects many facets of the collegiate experience, but is distinct enough to include as a separate attribute due to the level of commitment and investment required by student athletes. It is also interesting to note that varsity athletes at this large public institution benefit directly from higher education philanthropy, either through scholarships or program and facilities support. Therefore, possible motivations to give may include the desire to repay the institution for what was given to them.

The division collected two attributes relating to student athletes. The first feature, Athletic_Marker, denotes students that were involved in athletics at the club, commonwealth campus, or varsity level. As an additional descriptor, the VarsityAth_Marker denotes only Varsity athletes. It is necessary to include these as two separate attributes because there are numerous similarities between varsity and non-varsity athlete experiences, but varsity athletes enjoy scholarship and program support through philanthropy in addition to a more robust athletic experience.

Scholarship Recipients

Receiving a scholarship as a student may impact the student in a number of ways and demonstrate predictive power for later alumni giving. Need-based scholarship recipients may feel psychologically inclined to give back the grant money that enabled their educational experience. Merit-based scholarships may be a proxy for academic and future career success, and therefore a predictor of increased capacity to give. These anecdotal examples represent two of many reasons to explore the predictive power of scholarship award designation. Within this designation, the type of scholarship—need-based, academic merit-based, or extra-curricular—is not further specified or analyzed due to lack of data.

Active Alumni Association Membership

The only attribute included in this study that does not clearly describe alumni during their time as students is the marker for an active alumni membership (Active_Alum_Mem). The purpose of including this feature is twofold: (1) Many students gain a free or discounted membership in the Penn State Alumni Association either prior to graduation or shortly thereafter and (2) alumni association memberships, especially in the early years after graduation are almost exclusively driven by an affinity for the university forged during the student's experience. Therefore, while an alumnus may purchase the membership after his time as an undergraduate, an active alumni association membership is a proxy for a deep enough affinity for their experience to justify the costs of such membership.

Preprocessing and Filtering

Graduation Year Filter

Because giving varies significantly by age, which is likely a proxy for income and the capacity to give, limiting the study to a specified graduation year period is a key decision for this experiment. Recent alumni are much less likely to give immediately following graduation, so using the most recent data would include a strong negative bias for giving data.

Additionally, there is significant variability for recording of student data in the development database and little student information for graduates prior to 1991. The period from 2004-2009 represented the highest consecutive frequency of new entries and was governed by a consistent policy of recording those types of student attributes. Standard-age college graduates between 2009 and 2004 will be in the age range 27-32, with 6-10 years of opportunities to give.

Table 4: Student Activity Data Frequency

Year Added	Frequencies of Student Org Activity or Scholarship Recipient Entries	Percent of Total Entries
2004	14609	3.85
2005	16533	4.36
2006	15706	4.14
2007	16863	4.44
2008	16770	4.42
2009	16835	4.44

Exposure to Uniform Campus Experience

The large research higher education institution included in this study is structured as a distributed campus model. A large percentage of undergraduate students solely attend the “main” campus, while the remainder either complete some or all of their four year degree at one of 19 Commonwealth Campus

across the state of Pennsylvania. For the purposes of this study, the dataset was filtered to only include undergraduates who attended the University Park campus at some point and were therefore exposed to a reasonably uniform experience. Students who also attended a Commonwealth Campus in addition to University Park campus, are designated by the AttendCampus attribute. Because the majority of commonwealth campus students participate in a 2+2 program and graduate from University Park, a sizable portion of the instances share this designation.

Filtering by Giving Behavior

Within the Division of Development and Alumni Relations at this large public university, there are two categories of giving that are not characteristic or representative of academic philanthropy. While these two programs are not necessarily unique to this university, donors that gave exclusively to either program were treated as non-donors. Donors that had given at least one gift to another program were still included as donors. The two programs are THON, the aforementioned student-lead philanthropy program to combat childhood cancer and Athletics, which in some ways incentivizes gifts in exchange for football tickets. Neither program embodies the mission of development or academic philanthropy and represent specific affinities that are not considered typical higher education giving. Out of 15,549 donors from graduation years 2004-2009, 6,283 alumni were considered nondonors because of giving exclusively to athletics or THON.

Giving Statistics of Nominal Attributes

The nominal attributes each describe some facet of the student experience or affinity for the university. The distribution for status and number of donors are listed below.

Table 5: Counts of Donors By Nominal Attributes

Attribute	Y (Number of Donors)	N (Number of Donors)
Scholarship_Marker	11124 **(2161)**	31569 **(7103)**
Stud_Act_Marker	10187 **(3377)**	32506 **(5887)**
Athletic_Marker	2764 **(647)**	39929 **(8617)**
VarsityAth_Marker	875 **(245)**	41818 **(9020)**
Local_Student (Zip<15mi)	4090 **(596)**	38603 **(8669)**
Attend _Campus	36026 **(7947)**	6667 **(1318)**
First_Degree_PSU	27074 **(6855)**	15619 **(2409)**
Multi_PSU_Degrees	27074 **(936)**	15619 **(8328)**
Active_Alum_Member	5350 **(3325)**	37343 **(5939)**
PSU_DONOR	9264 **(9264)**	33429 **(0)**

For a dataset with 9264 donors, the attributes with the highest number of donors include: Attend_Campus, Active_Alum_Member, First_Degree_PSU, and Stud_Act_Marker. Attend_Campus denotes those who likely started at a commonwealth campus, but graduated from the University Park campus. Since students can come from 19 campuses to attend University Park, both the positive coding and donor values are large proportions of the whole. The remaining three features are in some ways linked by affinity for an aspect of the university. Active alumni membership denotes a desire to stay connected to the university; student activity participation implies an affinity for a certain student organization and investment in that

activity; and those attending Penn State as their first college will likely demonstrate an affinity for the university as their first collegiate experience.

Given a positive value for Stud_Act_Marker, the instance is further characterized by the type of student organization activity. Listed below is the topology of student organizations—coded manually with guidance from student affairs designations.

Table 6: Org Type Coding and Donor Distribution

Org Type	Percentage of all students coded	Percent Donors
Academic_Club	1%	24%
Academic_Honors_Society/Program	6%	23%
Campus_Leadership	3%	39%
Global_Program	0%	20%
Greek	29%	26%
Honors_Society	3%	50%
Mentoring_Program	4%	21%
Other	18%	29%
Penn_State_Pride_Organization	7%	41%
Performing_Arts	2%	19%
Philanthropic_Organization	5%	36%
Professional_Development_Club	1%	23%
Publishing/Communications_Club	1%	24%
Religious	0%	38%
ROTC	1%	20%
THON	17%	51%
No Student Activity Coding		18%

The wide distribution and relative sizes of the types of student organizations makes it difficult to make observations based solely on percentage of donors. It is also likely that not all students who participated in these activities are coded as participating. It is reasonable to assume that organization officers or those highly involved and willing to self-report largely comprise the set above.

The final nominal attribute of interest is college. While a higher aggregate level than major, college designations can provide important information regarding both student characteristics and capacity to give. The distribution of percent donors may also be influenced by the relative sizes and operations of each development unit within the colleges.

Table 7: Distribution of Alumni Giving by College

School Code	Percent Students Coded	Percent Donors
LIB	19%	24%
BUS	18%	27%
ENG	14%	21%
HHD	12%	25%
COM	10%	24%
SCI	7%	27%
EDU	6%	16%
AGR	5%	17%
A&A	4%	32%
IST	3%	19%
EMS	3%	21%
NUR	0%	30%
BPH	0%	41%
ALT	0%	5%
BRD	0%	8%
CAP	0%	0%
ABG	0%	0%
BLV	0%	0%
CWC	0%	0%

Chapter 3

Methodology

In addition to the richness and size of the data, the type of model has significant influence on the prediction accuracy. With the cost of running different models being relatively low, six different approaches will be employed for this study: Naïve Bayes, Decision Trees, Random Forest, AdaboostM1, Bagging, and Support Vector Machine algorithms. However, prior to running the model, a significant amount of effort is usually required to preprocess the data for machine learning use. This can include transforming variables, handling missing data, and filtering attributes. Some of the preprocessing work is dictated by the type of model if it requires a certain type of variables (nominal vs. numeric) or includes limitations with regards to handling missing values. These steps are explained below as the preprocessing and concessions made prior to running the model can have tangible effects on the results.

Weka

The machine learning client used in this study is Weka, or Waikato Environment for Knowledge Analysis, which is a popular suite of machine learning models coded in Java by the University of Waikato, New Zealand. The program is free to use under the General Public License and supports several data mining methods from preprocessing, to classification modeling, to visualization (Witten, Frank, & Hall, 2011).

Weka includes many useful tools for classification techniques and its classifiers are vetted and supported by the machine learning research community. The client provides a user-friendly GUI called the Explorer, which allows users to Preprocess, Classify, Cluster, Associate, Select Features, and

Visualize their data. While many of the tabs help users clean, explore, and visualize the data prior to modeling, the classify tab is classification engine of the client. Weka provides the following types of classifiers: Bayes, Functions, Lazy, Meta, MI, Rules, and Trees. Each classifier includes a description, default parameters, and a link to the source material.

After selecting the desired classifier, choosing the outcome class, and running the model, Weka provides several metrics for performance. These results help determine the most appropriate classifier for the dataset and the problem space, as well as inform the user on other classifiers or parameters to explore in order to improve performance. Brief definitions of these metrics are explained below:

Area Under Receiver Operating Characteristic (ROC) Curve

A curve, derived from the True Positive and False Positive rates that shows how robustly a classifier can distinguish positive and negative outcomes. An ideal ROC curve has 100% TP rate and 0% FP rate, whereas a positive, diagonal line, or 50%-50% split, represents a classifier that can classify instances with only 50% accuracy. The values of the curve are determined by varying the cutoff or threshold for classification. Varying the threshold, or sensitivity of the classifier will produce different quantities of true positive and negatives, and false positive and negatives. ROC curves visualize this indirect relationship and help uncover the ideal threshold for a classifier. The ROC metric listed after running a classifier is the Area under the ROC curve, which increases with performance between .5 and 1. ROC curves are particularly important to examine when the class distribution is unbalanced i.e. there are many more instances for one outcome compared to the others (Bradley, 1997).

Classification Accuracy

The first metric included in the summary, classification accuracy is defined as a percentage of correctly classified instances for the test set. While this may be a starting point for measuring performance, it cannot be the final determinant for whether or not a classifier is fitting the data correctly. For unbalanced datasets or unrepresentative training sets, a high classification accuracy can be misleading. For example, if 90% of instances are class A and 10% are class B, the model will be more effective at classifying instances because the probability of an instance belonging to class A is significantly higher than class B. Additionally, while cross-validation helps mitigate the effect of training set bias, an unrealistically similar training and test set pair will also skew the validity of a high classification accuracy. These caveats underscore the importance of also examining the other metrics available in Weka including: F-Measure and ROC Area.

Precision & Recall

Precision is the proportion of instances that truly have class A compared to the total number of instances classified as class A. Recall is the proportion of instances classified as class A compared to the actual total in that class (i.e. the True Positive rate). These two metrics have a somewhat unintuitive relationship. Depending on the sensitivity of the classifier, the precision and recall can fluctuate without necessarily signaling a better or worse classification performance. However, in cases when the goal is to maximize either TP rate or precision, these metrics can be used almost exclusively. Additionally, it is usually most helpful to measure the classifier on the Precision and Recall of the minority class in cases where the data is largely unbalanced.

F-Measure

F-measure is a combined measure for precision and recall, which helps demystify the tradeoffs inherent in an increase in either precision or recall and a subsequent decrease in the other metric. F-measure is calculated by:

$$Fmeasure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

F-measure values are usually compared across classifiers like Area under ROC (AUC) as a more reliable performance metric than classification accuracy and can range from 0 to 1, increasing with better classification.

Variable Transformation

For machine learning, the transformation or combination of variables is required in the preprocessing steps for many of the classifiers because the variables must be in a form that can be fit to a rule or logical test. The process of transforming or combining variables should be informed by background understanding of the problem and the dataset. An example of a simple transformation was changing a [0,1] binary coding into [Y,N] so that Weka did not mistake a "flag" attribute for a numeric attribute. An example of a more complex transformation is the creation of an additional attribute – OrgType – that clarified the Student_Act flag for those who participated in student organization. When variables clarify or are contingent on other binary variables there are many ways to handle the attribute representation. For this study, it was appropriate to include the organization type as a completely separate attribute because of the desire to examine and validate how different student activities are able to predict alumni giving.

Naïve Bayes Classifier

Naïve Bayes gets its name from Bayes' rule and works very effectively in certain circumstances despite the term naïve. Bayes' rule says that if you have a hypothesis H and evidence E relating to that hypothesis, then

$$P[H|E] = \frac{P[E|H] * P[H]}{P[E]}$$

When Bayes' rule is applied with the probabilities of outcome H given attributes E_1, E_2, … E_i, the result can be normalized and the outcome can be predicted for a new set of attributes. When combined with feature selection procedures that eliminate redundant attributes, Naïve Bayes performs very well on real datasets and is frequently used as a first step in understanding the predictive power of the dataset (Witten, Frank, & Hall, 2011).

While Naïve Bayes can handle unknown or missing values better than other models, the disadvantage of the approach stems from its theoretical basis (Michie, Spiegelhalter, & Taylor, 1994). Because of the derivation of the Bayes Rule, Naïve Bayes performs best if the features are conditionally independent for each class. For the student characteristic and alumni giving dataset, some features are conditionally independent (i.e. college and graduation year), while others are more closely conditionally related (i.e. student activity (y/n) and type of student activity (categorical)).

Recent comparisons to other classification techniques have found other methods to be superior to Naïve Bayes (Caruana & Niculescu-Mizil), but have explored the theoretical basis for why it performs better than expected given its operating assumptions (Zhang, 2004). Additionally, a benefit of Naïve Bayes is that, when appropriately used, it requires only a small amount of training data to be effective.

C4.5 (J48) Decision Trees Model

A decision tree uses tree-like graph decisions to learn a particular dataset, beginning with a single node and terminating each path at a predicted class through a series of decisions that test on the most relevant attributes. The concept of constructing a decision tree can be explained recursively: Select an attribute for the root node and make one branch for each possible value. The dataset can be sorted by that attribute along each of the branches. Repeating this process for each additional attribute, using only the instances that reach each additional branch, builds the full decision tree. It is easy to imagine the intractability problems in mapping out every possible path to an outcome when considering even a relatively small number of features. Constructing trees in full also reduces the interpretability and value of the results (Witten, Frank, & Hall, 2011). Therefore, the goal of decision tree learning is to construct a reasonably sized decision tree: small enough to be meaningful and manageable for the user, but complex enough to correctly fit the data.

There a variety of methods for building a decision tree—that is deciding which attribute to split on and how to "prune" the decision tree given a set of instances. One common concept for how to select the attribute for each node is known as information gain criterion. This process assigns an expected amount of information needed in order for a new instance to be classified. This metric is based on the purity of the instances that arrive at the node i.e. how uniform the classes are at that node (Witten, Frank, & Hall, 2011). With this information, choosing the next attribute to split on can be informed by the expected information gain of doing so. To minimize the size of the tree, one would maximize the information gain and allow classification to happen at the lower order nodes.

Once the tree is formed, there is also the potential issue of overfitting. Decision tree classifiers often include pruning processes to simplify the tree and ensure that it will be reasonably adaptable to variations in a new test set. The decision tree learning model employed for the purposes of this study is called J48, which is built on C4.5 design, and is based on information gain decision tree induction. J48 is particularly robust at dealing with numeric attributes, missing values and noisy data (Witten, Frank, &

Hall, 2011). J48 splits and selects attributes based on calculated information game and entropy values. Entropy is defined as the measurement of uncertainty in any random variable and ultimately used to generate the information gain according to the equation: $Gain(P, C) = Entropy(P - Entropy(C|P))$ (Bhargava, Bhargava, & Mathuria, 2013). Like most popular models, J48 also includes automatic pruning capabilities which further reduces the redundancy and unnecessary complexity of decision trees by consolidating and discontinuing certain branches that cause overfitting. One of the notable disadvantages of J48 is that the size of trees constructed in this way increase linearly with the number of instances. With a large number of examples and possible values per attribute, the tree may lose interpretability. (Bhargava, Bhargava, & Mathuria, 2013)

Random Forest

As an expansion of decision tree concepts, Random Forest creates many classification trees and aggregates the class results for a new instance—assigning the majority class from the entire forest. Each tree is grown using the following process:

1. Randomly sample N cases (where N is the size of the training set) with replacement from original data.
2. Randomly select a subset of attributes from the full set and best-split for each node down the tree.
3. Grow each tree to the largest extent possible.

During testing, instances are applied to each tree and a class is generated. Upon aggregation, the class with the highest number of "votes" is assigned to the instance. For this method, the error rate is determined by two factors: (1) the correlation between trees in the forest, where increased correlation results in increased error rate, and (2) the strength, or classification accuracy, of each tree in the forest. Generally, random forests are less likely than simple decision trees to over-fit and can also handle large

datasets or missing data gracefully. Increasing the number of trees in the forest will increase the performance of the model to a certain point, after which an increase in trees will plateau in performance.

AdaboostM1

AdaboostM1 is a boosting method that falls in the category of meta-classifiers because it employs another classifier with additional methods for reducing error and increasing performance. This goal is achieved by iteratively running a "weak" learning algorithm on different distributions of the training data and then combining the resulting classifiers into a single model. The improved performance stems from two core elements of boosting: (1) it generates a hypothesis with a lower error by combining hypotheses from each iteration over different test sets and (2) it reduces variance by taking a weighted sum of the training samples (Freund & Schapire, 1996).

Adaboost does well on datasets with one or both of the following properties: (1) instances have varying degrees of hardness and (2) the dataset causes significantly different hypotheses to be generated for different training sets. The latter can be achieved through selection of an underlying algorithm that has high sensitivity to changes in the data (Freund & Schapire, 1996). For this study, Decision Stump, a decision tree algorithm that selects a root node for classification by entropy calculations, is used as the selected base classifier. Additionally, a best practice for adaboost is to begin at 100 iterations and increase to ~200 in order to achieve maximum performance, after which increasing the number of iterations does not increase accuracy further.

In comparison with base classifiers performing better than chance, Adaboost performs significantly better than simple rules classifiers and marginally better than C.45 (or J48 in Weka). Compared to other meta classifiers, such as bagging, the results are generally similar in terms of performance (Freund & Schapire, 1996).

Bagging

Bagging (or **b**ootstrap **ag**gregat**ing**) is another meta classifier, which uses a base learner to generate multiple versions of a predictor to increase performance. Bagging loops over the different versions and does a "plurality" vote when predicting the outcome class. The dataset is randomly divided into a test set and learning set and a classification tree is constructed from the learning set. The learner is run on the test set and produces a misclassification (error) rate. Then a "bootstrap" sample is selected from the learning set and a tree is grown from the bootstrap sample. The original set L is used as a test set to select the best pruned tree. This process is repeat iteratively producing a classifier each time. In estimating the class of a new instance, the class with plurality among each of the bootstrap learners is selected (Breiman, 1996).

The critical determinant of improving accuracy with bagging is the stability of the procedure for constructing the base learner's predictor on each test set variation. If changes in the learning set produce small changes in the resulting predictor, performance will not increase dramatically. But with greater instability, bagging can make significant strides in classification accuracy. Therefore, it can amplify a good but unstable learner towards optimal performance, but will mirror or even slightly degrade a stable learner's performance (Breiman, 1996). Additionally, a best practice for bagging is to begin at 10 bootstrap iterations and increase ~100 to realize the maximum increase in performance.

Support Vector Machines (SVM) – Sequential Minimal Optimization (SMO)

A Support Vector Machine is a binary classifier with the discriminant function as a weighted sum of the kernel functions overall all training samples. In the simplest form, a SVM uses the kernel functions to create a hyperplane that separates the set of positive examples from the set of negative examples. For each test set, Lagranage multipliers are applied to solve a series of stepwise computational statements known as the Quadratic Programming (QP) problem (John C. Platt, 2000). These calculations are used for

optimization of the margins between the hyperplane and the nearest examples, which determines classification accuracy. The algorithm can be visualized at a high level by the figure below:

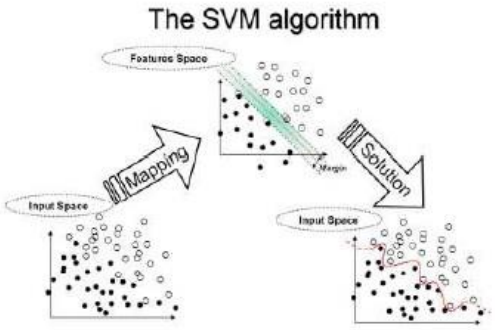

Figure 2: SVM Algorithm Flow – Source: dtreg.com

SMO is a unique approach to SVM because it chooses to solve the smallest possible optimization problem at every step and does not require iterative inner loops or matrix storage that usually characterize other support vector machines. Rather than executing expressions for the full QP problem, it selects the optimal sub-step to compute (John C. Platt, 2000). In this study, logistic modeling is enabled to offer a non-linear alternative to the other methods. Two different kernels are used, RBFKernal and PolyKernal, and the better performing result is included. This reduces execution duration and increases the maximum size of the dataset for which SMO is feasible. Generally, SMO is expected to perform better than simpler classifiers such as J48 or Naïve Bayes.

Parameter Adjustment

The parameters were generally set to default, but are described for replicability below:

Classifier Version Weka 3-6	Parameter Specification
Naïve Bayes	Default; 10 fold cross-validation
J48 Decision Tree	Default; 10 fold cross-validation

Random Forest (n=200)	Default; Num_iterations = 200; 10 fold cross-validation
AdaboostM1 (n=100)	Default; Num_iterations = 100; classifier = decision stump; 10 fold cross-validation
Bagging	Default; classifier = REPTree; 10 fold cross-validation
SMO	Default except Build Logistic Models TRUE; Kernel = PolyKernel

Figure 3: Parameters Specification

Chapter 4

Results and Discussion

Six classifiers were used to predict donor status of alumni using descriptive student data of undergraduates from class years 2004-2009 at a large, public research university. The data was imported into Weka and 10-fold cross validation was applied for each method. Generally, each classifier performed significantly better than chance (ROC Area = .500) in the binary classification of donor or non-donor, but there is little variation of performance between methods despite stark differences in conceptual theory and approach. The metrics for classification accuracy is listed for comparison below:

Table 8: Classifier Results

Classifier	Classification Accuracy	F-Measure (minority class)	ROC Area (minority class)	Number of Donors Classified Correctly
Naïve Bayes	81.36%	0.448	0.735	3222
J48 Decision Tree	81.17%	0.421	0.714	2923
Random Forest (n=200)	79.91%	0.403	0.694	2889
AdaboostM1 (n=100)	81.60%	0.444	0.736	3142
Bagging	81.30%	0.427	0.729	2976
SMO	81.35%	0.455	0.650	3325

For unbalanced datasets, i.e. class distributions are uneven, classification accuracy is often not the best metric to measure performance. A high classification accuracy may indicate the ability to identify one class, but not the other. In this case, where the minority class is the class of interest (PSU_DONOR = Y), AUR, or Area under ROC or F-measure are more appropriate for

benchmarking performance. Based on the model outputs, AdaboostM1, a meta-classifier, had the highest performance by AUR, correctly classifying 3142 donors. However, this level of accuracy was nearly matched by Naïve Bayes, a simpler yet surprisingly effective classifier. Bagging and J48 Decision Trees had comparable AUR values, but slightly lower F-Measure and True Positive (TP) classifications. This is intuitive as Bagging employs a fast tree decision learner, REPTree, as a base classifier and improves on those results through bootstrapping.

Random Forest and SMO produced somewhat surprising results. Typically, Random Forest outperforms J48 and simpler learning classifiers, especially with the increasing of trees built, up to a certain point. However, it bears the lowest F-measure and True Positive rate and clearly under performs NaiveBayes and J48. One possible reason for the discrepancy is that unbalanced class distributions can affect RandomForest more than other classifiers, especially in prediction of the minority class. Additionally, it is also possible that RandomForest is overfitting the data, likely due to the divergent attributes with many options such as Pref_School_Code and Org_Type. SMO, a support vector machine classifier, also has interesting results. While it has a low F-Measure and AUR, it correctly classifies the most number of donors. This may underscore an important distinction of SMO. While additional factors and noisy data degrade performance for decision tree classifiers, SMO is more immune to these challenges due to its ability to select a subset of instances (support vector) that defines the class. The type of feature selection, increases the runtime and complexity, but reduces the susceptibility to misleading or irrelevant correlations between variables.

Ranking and Selecting Attributes

One of the inherent challenges of large datasets and classification problems is determining which attributes are relevant for prediction and information-rich. Features that are either redundant or hold little predictive power add unnecessary complexity to the model and reduce the interpretability of the results. However, because of the potential hidden relationships between attributes, it is beneficial to refrain from removing attributes based on intuition. To improve machine learning algorithms and mitigate computation challenges of high dimensionality, the subfield of feature selection has emerged to supplement learning classifiers. These algorithms operate in a similar train-test process over the dataset and apply one of several different evaluators in order to determine rank of predictive power for the given attributes and output class.

Weka includes a tab dedicated specifically to feature selection, housing several ranking algorithms. For unbalanced class data, it is helpful to use the Gain Ratio Attribute Evaluator (Novakovic, 2009) with the Ranker search method because this particular evaluator normalizes the information gain ratio by class according to the following expression: GainR(Class, Attribute) = (H(Class) - H(Class | Attribute)) / H(Attribute). As explained in the decision tree methodology, information gain is a metric to quantify how much a given attribute contributes to being able to achieve maximum purity of class. The attribute rankings for dataset is listed below:

Table 9: Attribute Ranking by Information Gain Ratio

Rank	Information Gain Ratio	Attribute
1	0.150439	Active_Alum_Member
2	0.020588	Stud_Act_Marker
3	0.010581	First_Degree_PSU
4	0.005452	Local_Student (Zip<15mi)

5	0.003451	Degree_Year
6	0.002221	VarsityAth_Marker
7	0.001666	Multi_PSU_Degrees
8	0.00106	Pref_School_Code
9	0.00095	Scholarship_Marker
10	0.000833	Org_Type
11	0.000484	Attend_Campus
12	0.000244	Athletic_Marker
13	0.000111	Gender

The Active_Alum_Member attribute, denoting membership in the university's alumni association, is clearly the most relevant feature for determining alumni giving. This finding is logical both intuitively and based on the high percentage of alumni with memberships that are also donors (62%). It is also interesting that participation in a student activity (Stud_Act_Marker) has a high predictive power, yet the type of participation (Org_Type) is ranked significantly lower. Perhaps because students feel a strong affinity for their first degree program, First_PSU_Degree is a strong determinant for classification as well.

The ranker also demonstrates that gender, athletic participation, scholarship status, and even college designation do not provide high predictive power. These attributes neither share similar distributions nor exhibit an obvious cross-correlation. Therefore, they just may not be ideal characteristics to include in modeling as they likely add little value. Further, by including Org_Type and Pref_School_Code, the complexity of the model increases significantly due to the number of potential nominal values.

To determine if the majority of the predictive power stemmed from the top five attributes, the J48 classifier was run on ranked features 1-5 and compared to the full model. The results are summarized below:

Attribute Set	Classification Accuracy	F-Measure	Area under ROC	Correctly Classified Donors	Tree Size	Number of Leaves
Top 5 by Ranker	81.48%	0.421	0.65	2869	17	9
Full Set (13)	81.17%	0.421	.714	2923	1177	957

The feature selected decision tree is smaller by a factor of 70 with similar classification accuracy and slightly lower AUR (Area under ROC). For practical purposes, it would correctly classify 54 less donors, but significantly increase interpretability of the results for use by a development professional. It seems to hold that the predictive power largely lies with the top five attributes and that the two divergent attributes (Org_Type and Pref_School_Code) are responsible for the drastic increase in tree size and number of leaves.

Simple Logistic Regression Modeling

To generate a comparison to other types of analysis, a simple logistic regression was applied to the dataset to determine the relative coefficients and classification performance of a traditional method. The weighted coefficients used to build the expression for classification deviated significantly from both the feature selection by gain ratio and decision tree higher order nodes.

Table 10: SimpleLogistic Regression Coefficients

Attribute (=Value)	Coefficient
[Scholarship_Marker]	0.08
[Org_Type=Honors_Society]	0.35
[Org_Type=Philanthropic_Organization]	0.23
[Org_Type=THON]	0.43
[Org_Type=Penn_State_Pride_Organization]	0.13
[Org_Type=Professional_Development_Club]	-0.32
[Org_Type=ROTC]	-0.33
[Stud_Act_Marker]	-0.26
[VarsityAth_Marker]	0.14

[Degree_Year]	-0.07
[Pref_School_Code=LIB]	-0.05
[Pref_School_Code=ENG]	0.07
[Pref_School_Code=BUS]	0.09
[Pref_School_Code=HHD]	-0.06
[Pref_School_Code=SCI]	-0.07
[Local_Student(Zip<15mi)]	-0.17
[Attend_Campus]	-0.1
[First_Degree_PSU]	0.21
[Active_Alum_Member]	1.03
	Constant = 143.47

This shows that logistic regression, an entrenched method for data mining, can be a helpful starting point and alternative method for assessing predictive power of different attributes. While the feature selection did not find significant predictive power in Pref_School_Code and variations in Org_Type, these features appear several times in the logistic regression. Additionally, it is important to note that the theoretical mechanism of logistic regression is similar to Naïve Bayes, which may result in similar performance when run on the same dataset.

The classification accuracy is characterized by an ROC of .737, an F-Measure of .44, and a correct classification rate of 81.6%. These results suggest a fairly robust process in determining the class, but rely largely on itemizing specific subpopulations with very specific features – either by major or organization type. In this case, logistic regression may perform well on this specific dataset but fail to capture relationships between features in larger and higher feature count datasets.

Chapter 5

Conclusions & Implications

Problems with complex relationships between attributes and a practical need for accurate classification have much to gain from machine learning classifiers. In the space of philanthropic giving, the complexity of the decision makes it difficult for fundraisers to filter, narrow, and target specific subpopulations to increase their productivity. This problem become even more complex when considering universities, a pseudo-philanthropic organization that had tangible impact on the potential donor. This study, among other things, confirms that there is merit in (1) examining characteristics that describe the student experience for predictive purposes and (2) that an approach that considers the full attribute set in a robust way improves classification accuracy and interpretability of results.

In the first seven years of alumni status, an alumnus can receive up to 25 solicitations either as a mass appeal that all graduates of that year received or because of filtering by some surface demographic attribute i.e. college, major, gender. The monetary cost of "catch-all," sweeping solicitation campaigns are significant, but the absence of modeling also causes another cost for the organization. Ad hoc or demographic based audience filtering results in generic messages that may or may not resonate with recipients. Using the classifiers outlined in this study, particularly the ones that offer a reasonable level of interpretability, arms development professionals with richer information to refine messaging and strategy for a greatly reduced subset of likely donors.

Based on the results of 6 different classifiers, the included attributes have enough predictive power to build a viable alumni giving prediction model. Each model performs significantly better than chance with AUR values of approximately .7 and correct classification

of ~3200 donors out of a population of ~42,000 young alumni. The percentage of alumni in this dataset that made a gift was 27%. The simplest method for capturing alumni donors with this information would be a blind solicitation effort, biased at p=.27. Applied to this group of alumni, only 2500 of the 11,349 alumni solicited would go on to make a gift – a 22% donation rate. If each mailer costs 30 cents and the average gift of $100.00, this method would cost $3,400 and generate $250,000. However, using the classification modeling to target the 3200 donors selected would cost $960 and generate 320,000. In addition to the 28% increase in revenue and a decrease in cost by a factor of 3.5, this solicitation method would enable data-driven segmentation—delivering tailored messaging rather than generic appeals. Tailored messages generally improve return rate and average donation, which would further increase the gap between the blind solicitation and the machine learning approaches. As an initial study into the predictive power of this type of data and the value of ML methods, these results strongly support the advantage of classification models in identifying donors, enabling focused appeals, and increasing efficiency of solicitation efforts.

In choosing the most appropriate learning classifier for this problem set, it is important to balance two perhaps competing priorities: classification accuracy and interpretability of the results. Of course, driving up classification accuracy for the positive case is important, but these results must be operationalized by development professionals. Understanding the features at play in identifying likely donors may determine solicitation strategy, priority of the donor, and gift allocation. Bagging, J48, and Random Forest are decision tree based. In ideal cases, decision trees are intuitive and split along values that make sense to the user. In this case, the decision tree is cluttered by complexity and a large number of leaves. While it does provide a visual representation of the algorithm classification process, the size may be initially overwhelming. A

similar conclusion can be drawn for AdaboostM1, a high-performing algorithm with a lengthy and complex classification output. SMO provides a slightly more interpretable output and Naïve Bayes offers a straightforward frequency count of class assignments by attribute. Coupled with the performance metrics discussed in Chapter 4, the interpretability comparison suggests that Naïve Bayes or SMO would be the best models for this application.

The relative consistency for performance across classifiers suggests that the resulting accuracy was perhaps near the maximum for this particular dataset. When different methods produce different results, it is usually a signal to the user that performance can be optimized by tweaking parameters or exploring different classifiers. The varied set of algorithms hovering within a few decimal places in classification accuracy shows that this is not the case for this dataset. It is also important to note that the set of classifiers includes both linear and nonlinear classifiers. While it helpful to begin with a linear classifier, poor performance can motivate users to see if a nonlinear model will better fit the data. To try and optimize performance, both decision trees and service vector machines were included as the nonlinear methods. Decision trees are particularly good at describing nonlinear relationships with low feature counts, but in this case, did not significantly outperform linear classifiers. SMO can be configured with a nonlinear polynomial kernel and achieved a slightly higher correct classification number, but again not high enough to suggest the data was too nonlinear for classifiers such as Naïve Bayes.

Suggestions for Improvement

As an initial exploration of the link between student-centered data and alumni giving, there are several suggestions for improvement and further research. The first revolves around the data collection and scope and the second involves more extensive modeling.

The data used in this study represents a small fraction of the total student information already captured by universities. Decentralization of data collection and limits of use for that data hinders the inclusion of all relevant information by development analysts. For example, GPA is housed in academic units, while richer scholarship information is collected and restricted by admissions for financial privacy. It is easy to imagine the many places one could search for relevant student information that could be useful in modeling. Based off of the projected increase in identification and revenue from targeted modeling, this study supports the broader and deeper collection of information by the division as the primary way to increase performance. Examples could include: housing information, attendance at sporting or programming events, majors and minors, awards received, internships, disciplinary information, academic performance, number of relatives that also attended, etc. Similar processes of filtering and transforming these variables would be necessary steps for integration.

Besides supplementing the types and richness of the data, the research questions could be broadened significantly. While this study tested donor status, similar classifiers could be applied to evaluate levels of giving. Even further, the frequency of giving or the time to first gift could be modeled using these techniques. Giving patterns are complex and warrant evaluation on multiple levels. This data presents enough evidence for predictive power that it may be worth using for a slightly different research question related to alumni giving.

Finally, since the modeled attributes are now collected by the division for every graduating class, this model can be expanded to recent alums beyond 2009 and even modified to predict donors in the most recent graduating classes. Building a pipeline of young alumni donors, using data that has been available but not previously utilized, could provide tremendous value for the organization both in the receipt of immediate gifts and in the cultivation of long-term donors.

As attributes and graduation classes are added, this study can be replicated to determine the ideal classifier for both accuracy and interpretability. It represents an initial but sustainable method for identifying likely alumni donors year over year for improved efficiency and a more targeted approach.

BIBLIOGRAPHY

Bekkers, R., & Wiepking, P. (2007). *A Literature of Empiracal Studies of Philanthropy: Eight Mechanisms that Drive Charitable Giving.* Amsterdam, Netherlands: Center for Philanthropic Studies.

Bhargava, N., Bhargava, R., & Mathuria, M. (2013). Decision Tree Analysis on J48 Algorithm for Data Mining. *Intrenational Journal of Advanced Research in Computer Science and Software Engineering*.

Bradley, A. (1997). The Use of Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition*, 1145-1159.

Breiman. (1996). Bagging Predictors. *Machine Learning*, 123-140.

Carolina Angel. (2015, March 22). *Fundraising Fundamentals*. Retrieved from Council of Advancement and Support of Education: http://www.case.org/Publications_and_Products/Fundraising_Fundamentals_Intro/Fundraising_Fundamentals_section_7/Fundraising_Fundamentals_section_73.html

Caruana, R., & Niculescu-Mizil, A. (n.d.). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd international Conference on Machine Learning*, (p. 2006).

Caruthers, F. (1971). *A study of certain characteristics of alumni who provide financial support and alumni who provide no financial support for their alma mater.* Stillwater, OK: Unpublished Doctoral Dissertation.

Freund, Y., & Schapire, R. E. (1996). Experiments with a New Boosting Algorithm . *Machine Learning: Proceedings of the Thirteenth International Conference*.

Gaier, S. (2005). Alumni Satisfaction with Their Undergraduate Academic Experience and the Impact on Alumni Giving and Participation. *International Journal of Educational Advancement*, 279-288.

Hall, J. O. (1967). *A survey of the attitudes of Cornell alumni toward their alma mater.* Ithica, NY: Unpublished doctoral dissertation.

John C. Platt. (2000). *Fast Training of Support Vector Machines using Sequential Minimal Optimization.* Redmond, WA: Microsoft Research.

Key, J. (2001). Enhancing Fundraising Success with Custom Data Modeling. *International Journal of Nonprpofit and Voluntary Sector Marketing*, 335-346.

Korvas, R. J. (1984). *The relationship of selected alumni characteristics and attitudes to alumni financial support at private colleges.* Kansas City: University of Missouri at Kansas City.

Lara, C., & Johnson, D. K. (2008). *The Anatomy of a Likely Donor: Econometric Evidence on Philanthropy to Higher Education.* Colorado Springs, Colorado: Department of Economics and Business Colorado College.

Lindahl, W. E., & Winship, C. (1992). Predictive Models for Annual Fundraising and Major Gift Fundraising. *Nonprofit Management and Leadership*, 43-64.

Lindahl, W. E., & Winship, C. (1994). A Logit Model with Interactions for Predicting Major Gift Donors. *Research in Higher Education*, 729-743.

Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (1994). *Machine Learning, Neural and Statistical Classification.*

Miller, M. T., & Casebeer, A. L. (1991). *Donor Characteristics of College of Education Alumni: Examining Undergraduate Involvement.* U.S. Department of Education - Office of Education Research and Improvement.

Mingers, J. (1989). An Empirical Comparison of Pruning Methods. *Machine Learning*, 227-243.

Monks, J. (2002). Patterns of giving to one' salma mater among young graduates from selective institutions. *Economics of Education Review*, 121-130.

Novakovic, J. (2009). Using Information Gain Attribute Evaluation to Classify Sonar Targets. *17th Telecommunications forum* .

O'Connor, W. (1961). A study of certain factors characteristic of alumni who provide financial support and alumni who provide no financial support for their college. *Unpublished doctoral dissertation, University of Buffalo*.

Radcliffe, S. (2011). *A Study of Alumni Engagement and Its Relationship to Giving Behaviors* . Bucknell Master's Theses.

Russell, S., & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach (Third Edition).* Prentice Hall.

Taylor, A. L., & Martin, J. C. (1995). Characteristics of Alumni Donors and Nondonors at a Research I, Public University. *Research in Higher Education*, 283-301.

Thomas, J. A., & Smart, J. (2005). The Relationship between Personal and Social Growth and Involvement in College and Subsequent Alumni Giving. *Association for Institutional Research*.

Weerts, D. J., & Ronca, J. M. (2009). Using Classification Trees to Predict Alumni Giving for Higher Education. *Education Economics*, 95-122.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques.* Burlington, MA: Elsevier Inc.

Zhang, H. (2004). The Optimality of Naive Bayes. *FLAIRS2004 conference.* http://www.cs.unb.ca/profs/hzhang/.

ACADEMIC **VITA**

Dominic Mirabile
305 E. Prospect Ave. State College PA 16801
Dsm5265@psu.edu

EDUCATION:

THE PENNSYLVANIA STATE UNIVERSITY: Schreyer Honors College, College of Engineering—University Park, PA
Bachelor of Science in Electrical Engineering, May 2015
Minor: Engineering Leadership Development

ACADEMIC HIGHLIGHTS:

Dean's List: Fall 2011 through Spring 2013
President's Medal in recognition of 4.0 GPA – Fall 2012
Madden Engineering Scholarship, Gabron Electrical Engineering Scholarship, Schreyer Honors Scholarship
SAT: Math-800 | Reading-730

PROFESSIONAL EXPERIENCE:

NATIONAL INSTRUMENTS: Engineering Leadership Development—Austin, TX
Technical Marketing Intern, Summer 2014

- Demonstrated business and technical leadership while interfacing with customers, generating marketing materials, and executing department projects.
- Composed product training manuals, case studies, presentations, and value propositions that strengthened product branding in a targeted market space.
- Worked with department leads to brainstorm ways to better utilize big data, R programming, and web analytics for department success.

VERIZON WIRELESS: Network System Performance Division—Plymouth Meeting, PA
Electrical Engineering Intern, Summer 2013

- Actively monitored Philadelphia Tri-State Area's Network, comprised of thousands of RF cells, 6 major access sites, and 2 distribution hubs.
- Performed daily data analysis on link latencies, loading, and network health for resilient outage response and data speed optimization.
- Configured a high-level, real-time analysis tool that integrated multiple data streams into a simple platform for more robust monitoring and diagnostic functionality.

LEADERSHIP EXPERIENCE:

PENN STATE DANCE MARATHON (THON): Director of Donor & Alumni Relations
(March 2013—March 2014)

- Directly coordinated the fundraising efforts of the largest student-led philanthropy in the world, which raised a record-breaking $13,343,500 for pediatric cancer care and research—an 8% increase from the 2013 total
- Responsible for fundraising-related executive decisions by utilizing analysis of donor demographics, giving trends, and historical data
- Drove a 36% increase in corporate revenue to $1.75M by establishing and leveraging professional relationships with 2,900 companies, including several major gift donors
- Restructured committee direction to maximize alumni engagement by organizing a targeted campaign that reached 200 alumni groups and 70,000 households
- Oversaw planning and execution of a 46-hour dance marathon with an estimated 50,000 spectators, 3,800 volunteers, and 700 dancers

SPARK – PENN STATE'S ANALYTICS THINK TANK: Co-Founder & President

- Founded an open organization to bridge the gap between curriculum and data science careers by engaging students in big data projects, discussions, seminars, and development opportunities.
- Currently manage marketing campaigns, project development, and administrative tasks.
- Interface with relevant companies, alumni, and faculty to elicit buy-in and create value for the organization

RELEVANT SKILLS:

Business: Fundraising, Competitive analysis, Management, Marketing and Branding, AGILE Project Management, and Negotiation
Technical Skills: Proficient in Excel, R Programming, LabVIEW, Microsoft Word, Powerpoint, and Technical Writing. Experienced in Java, C++, MatLAB, and HTML.
Learning: Machine Learning, Algorithm Development, SAS