Finding a Minimum Variable Set to Predict College Enrollment

Grace Lee
Statistics and Data Science
Carnegie Mellon University
ghlee@andrew.cmu.edu

Andrew Resnikoff
Statistics and Data Science
Carnegie Mellon University
aresniko@andrew.cmu.edu

Leo Yoon
Statistics and Data Science
Carnegie Mellon University
sungjuny@andrew.cmu.edu

Abstract

The client, a predictive analytics company called [CLIENT OMITTED], is looking to better service educational institutions that are interested in predicting enrollment for prospective students but have only a limited amount of data available. The purpose of this investigation is to determine whether there exists a Minimum Variable Set that can be used to predict college enrollment for a generic school. Exploratory data analysis showed that it was possible to partition the full variable set into four categories: academic ability, demographics, interest and financial aid. A proposed stochastic variable selection method successfully identified 17 variables that spanned across the four categories. For each of [CLIENT OMITTED]'s existing customers, these variables were able to outperform naive enrollment predictions for admitted students using average Area Under the Receiver Operating Curve as a comparison metric.

Contents

1 Introduction

Each year, approximately five million students apply to one or more of the 5,300 colleges within the United States. These colleges are then tasked with deciding which applicants to admit, based on a variety of factors such as academic standards and personal essays. The difficulty for these institutions is that with each set of admission decisions brings the risk that too many or too few students will actually enroll. If too many students choose to enroll, the college will not have enough resources to provide an adequate education experience for the incoming class without significant financial investment. If too few students choose to enroll, the college is not able to raise enough revenue to cover operational expenses such as professors salaries. In either scenario, an institution faces potentially devastating financial impact and students do not receive an optimal education experience. Thus, it is crucial that colleges can accurately assess the likelihood that a given admitted student will enroll so that they can make responsible admission decisions.

[CLIENT OMITTED] is a predictive analytics company that focuses on answering High-Impact Questions (HIQs) in different fields. The college enrollment problem is one such HIQ that [CLIENT OMITTED] has developed its own proprietary solution for. In their models, they rely on large sets of covariates provided by colleges to predict the likelihood that a given student will enroll. Unfortunately, collecting data is expensive; therefore, it is beneficial to have a model that is able to adequately predict enrollment with a smaller set of variables. Thus, the purpose of this investigation is to find the smallest set of variables that can be used to effectively estimate the probability of a given student's enrollment. This way, a potential new [CLIENT OMITTED] client with limited funds could collect information based on this variable set and still obtain accurate enrollment predictions.

Our main question of interest is finding the minimum variable set that potential new clients with limited data can utilize, which we will do using data on students collected by 10 different universities.

To determine this Minimum Variable Set, [CLIENT OMITTED] has provided a collection of 13 datasets. Each dataset represents the pool of prospective students who are of interest to a given college, with four of those belonging to one college with multiple campuses. Each observation contains information about a specific potential student collected by the college that relates to academics, demographics or other factors that the college deemed worthy of collecting. The contents of each dataset vary in both the number of potential students as well as the variables that are available for each student. Combined, the number of unique columns across the datasets is slightly greater than 500, with approximately 100 variables per school. Some schools collect data on only a few thousand prospective students, while others have information on over 100,000 prospective students. The response variable of interest is a binary indicator for whether or not the student eventually enrolled in the college.

2 Methods

2.1 Exploratory Data Analysis

2.1.1 Data Exploration

Across all datasets, there are approximately 500 unique variables. A histogram of variable frequencies in each of the 13 datasets is shown in Figure 1. Notice that the histogram is extremely right skewed and that a majority of variables are not available in most datasets. This is due both to different naming conventions across schools as well as differences in the types of information each school collects. This motivates the grouping of variables into categories that are uniform across all customers.

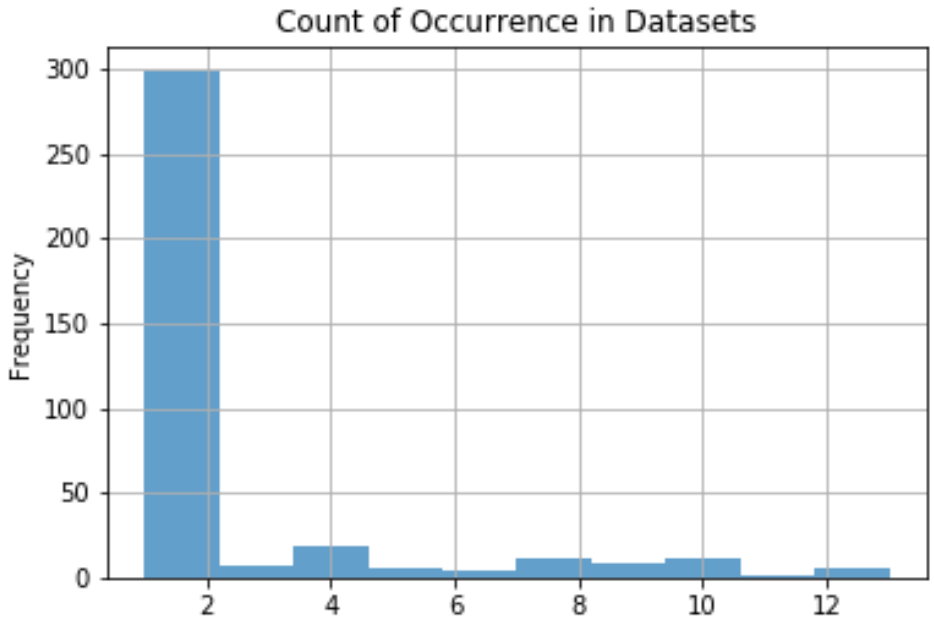


Figure 1: Histogram of Variable Occurrence Frequencies

Four major categories are able to encompass each of the variables while still explaining college enrollment: Academics, Demographics, Interest and Financial Aid. An example grouping is shown in Table 1.

Academics	**Demographics**	**Interest**	**Financial Aid**
High School GPA	Sex	Legacy Status	Expected Family Contribution
SAT Score	Race	Campus Visits	Scholarships Received
High School Rank	State	Athlete	Government Aid Received

Table 1: Examples of grouping of variables into the four major categories

The academics category includes fields such as students grade point average (GPA), SAT score, ACT score and percentile rank among their high school class. Academics should be predictive of enrollment in that we should expect those that are underqualified to not be admitted and those that are overqualified to choose a more competitive school. Histograms of GPA for accepted students by enrollment status in Figure 2 indicate that once a student is accepted, there is not a noticeable difference in GPA between the enrolled and not enrolled populations. Similar results are seen in plots of SAT and ACT standardized test scores. This finding is fairly intuitive, as we would expect that academics would decide where a given student might apply and be accepted but other factors would lead a student to choose between two institutions of similar academic reputation.

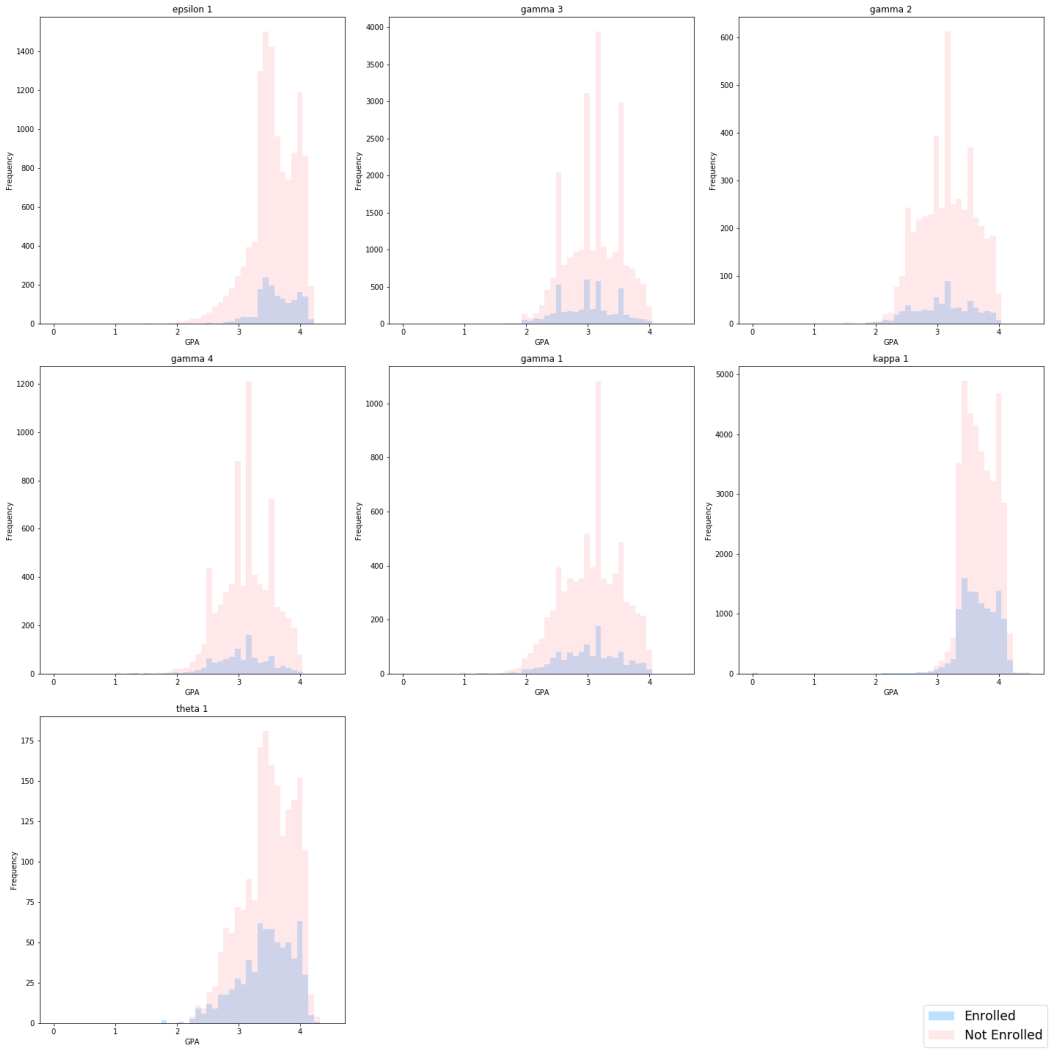
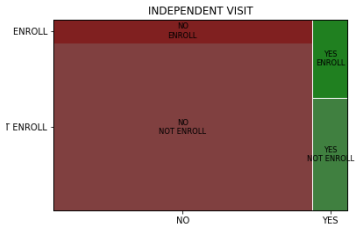


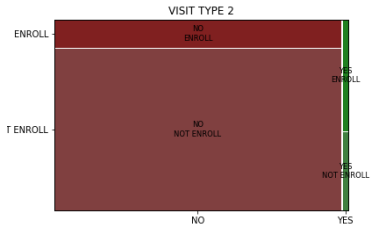
Figure 2: Histograms of GPA by Enrollment Status

Demographic factors of interest range from immutable factors such as ethnicity and gender to location based information like median income level and nationality. The distribution of ethnicities and gender varies from school to school in terms of enrollment populations and in general most demographic variables ultimately do not appear to be predictive by themselves. However, there is a strong possibility that some of these factors may interact with variables from other categories such as proxy variables for socioeconomic status with financial aid awards.

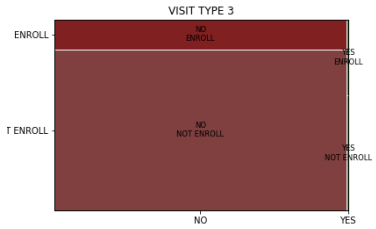
Interest level variables include binary indicators for school specific events and visits as well as other factors that may make a student more likely to be interested in attending such as legacy status. Mosaic plots in Figure 3 demonstrate that, at Gamma 1, some types of visits are associated with enrollment at higher rates than others. Histograms show that among enrolled students, legacy status students possess the same academic level of academic achievement as non legacy status students. This finding was contrary to the initial expectation that legacy students with lower standardized scores may be more likely to attend schools that are willing to accept them because of their legacy status.



(a) Independent Visit



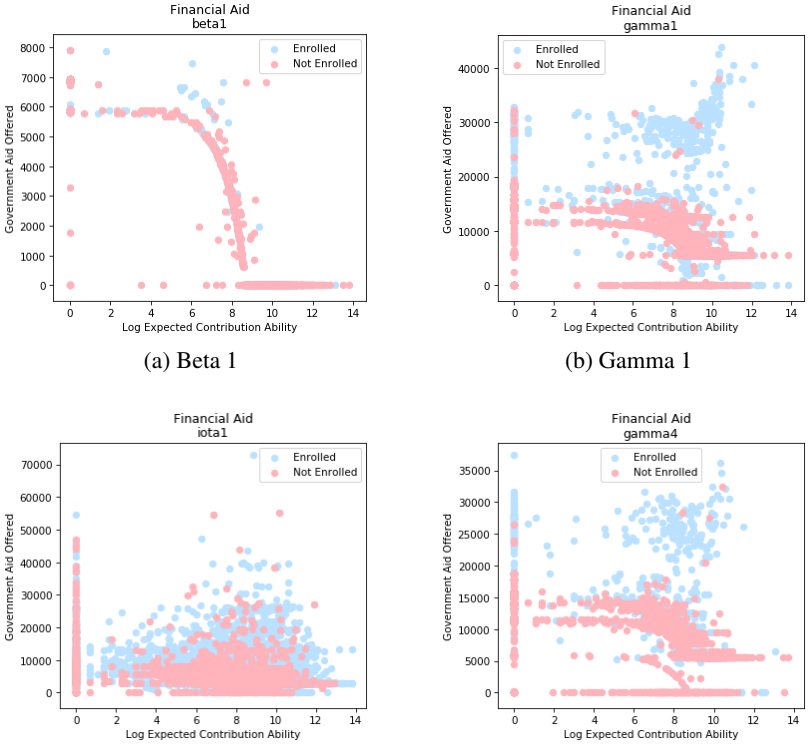
(b) Visit Type 2



(c) Visit Type 3

Figure 3: Mosaic Plots of Visits for Gamma 1

The category that we expect to have the most impact on our predictions is financial aid. Ultimately, students will only attend schools that they can afford to go to. Receiving aid or scholarships that can lower the cost burden on a students family can weigh heavily on their decision of which school to enroll. Scatterplots in Figure 4 exemplify this phenomenon, showing a dramatic separation between enrolled and not enrolled populations (among accepted students) for a given measure of how much a family can contribute financially out of pocket to a students education costs based on the amount of government financial assistance offered.



(a) Beta 1

(b) Gamma 1

(c) Iota 1

(d) Gamma 4

Figure 4: Plots of Government Aid offered against log Contribution Ability by Enrollment Status

Variable importance plots are constructed by fitting random forests to each dataset using all available predictors to classify enrollment status. Examples of these plots for customers Beta 1 and Gamma 3 are displayed in Figure 5. Notice that across schools, there are no variables that appear to be the most important in predicting enrollment status. Also, within a given school, none of the supposed overarching categories emerge as being solely responsible for predicting student enrollment. Thus, it is evident that an effective minimum variable set should include variables from each of the four categories in order to capture the differences in variable importance across schools.

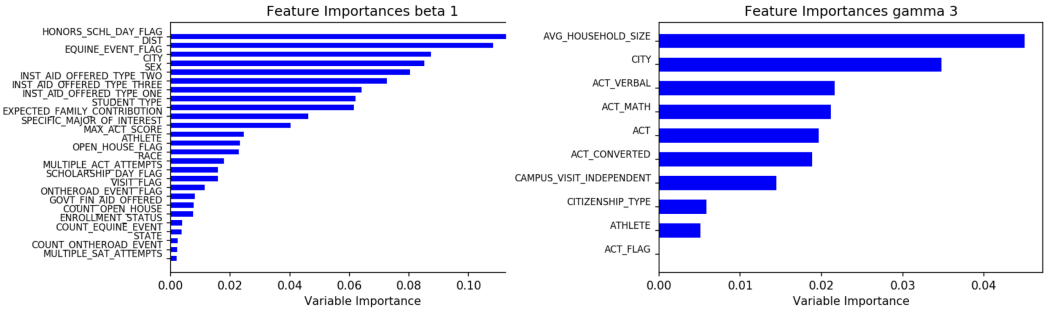


Figure 5: Variable Importance Plots Using Full Variable Set for Beta 1 and Gamma 3

2.1.2 Enrollment Lifecycles

A major feature of the enrollment process for colleges is that it happens in stages, hereby referred to as “lifecycles”. A student might become known to a school as a prospect or could apply to the school without the school never knowing about them. These lifecycles vary greatly from school to school,

although each school does have an admitted stage where most information about the students in that stage is available to the customer. Thus, we focus the investigation on determining the Minimum Variable Set to predict enrollment for admitted students.

2.1.3 Missing Data

Even after limiting observations to the admitted lifecycle, where the most amount of information is available, there still exists structural and non structural missingness throughout the data. For example, high proportions of missing standardized test data for Gamma 1 as shown in Table 2 may be indicative that Gamma 1 does not value standardized test scores in admission decisions and that the missingness is structural. In fact, Gamma 1 does not require students to send their scores to apply. This makes the population of students who have test scores different from those students who don't have test scores. χ^2 tests for independence confirm that in as many as seven of the schools, test scores are missing not at random. In the case of Alpha 1, a school that does consider standardized test scores, again there is missingness within the data. Table 3 shows that while ACT and SAT scores have substantial missingness, almost 99% of admitted students submitted at least one of ACT and SAT scores. Since the choice between taking ACT and SAT scores is usually decided by a student's geographic location, this missingness may be correlated with variables such as State or Distance.

Variable	**Proportion of Missing Values**
SAT Math Score	94.53%
SAT Reading Score	94.46%
SAT Total Score	94.45%
Converted ACT Score	93.48%
ACT Score	92.75%
Max Test Score	88.45%
ACT Verbal Score	84.34%
ACT Math Score	84.33%

Table 2: Proportion of Standardized Test Scores Missing for Gamma 1

Variable	**Proportion of Missing Values**
ACT Score	16.46%
SAT Score	69.41%
Max Test Score	1.82%

Table 3: Proportion of Standardized Test Scores Missing for Alpha 1

This large volume of missingness will have to be handled in different ways. Since the goal is to build a Minimum Variable Set that explains the data for all customers, it is reasonable to exclude variables that only have data available for select students (or penalize the variable if it exists across schools). However, it may be necessary to construct new columns to ensure important information is not lost.

2.2 Data Cleaning

2.2.1 Variable Reduction

Considering all of the given datasets, there are hundreds of unique variables that need to be narrowed down into a minimum variable set that we could interpret. Intuition suggests that a minimum vari-

able set should prefer to include variables that are shared among multiple schools. A plot of prediction accuracy against variable index sorted by frequency of occurrence in the collection of datasets shown in Figure 6 demonstrates that a cutoff of four schools is able to remove a large proportion of variables from the variable set while still keeping variables that provide accuracy throughout. Thus, by removing variables that do not appear in at least four datasets, we can reduce the total number of considered variables by over 80% while still maintaining most of the prediction accuracy. The white space in the plot represents variables that were not available in the admitted lifecycle.

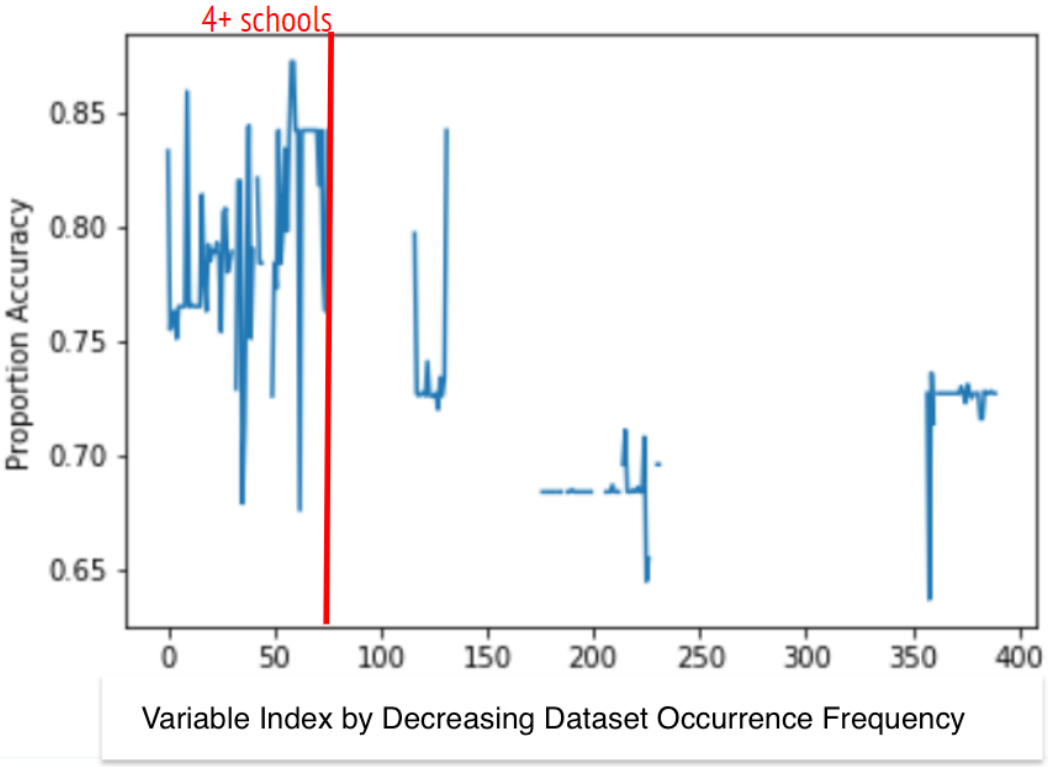


Figure 6: Logistic Regression Accuracy by Variable, Sorted by Frequency of Dataset Occurrence

2.2.2 Variable Consolidation

Since the way information is represented is not consistent from one school to the next, it is necessary to rename and consolidate variables so that each school's dataset is as similar to the others as possible.

The most obvious variables to be processed were academic variables related to standardized test scores. Since ACT score and SAT score are highly correlated, we can simply convert ACT scores to SAT scores[1] on the 1600 scale and take the greater of the two if both exist for the same student. This metric should be a suitable estimate for a student's standardized test performance if they submitted scores to the school in which they were accepted. Note that this does not improve missingness for schools where standardized tests were optional, such as Gamma.

Interest level data consisted of two major factors: campus visits and school events attended. Each school had some subset of these binary indicator variables, where each variable in the subset represented a specific type of visit or event for that school. These factors were transformed using a method hereby referred to as *Relative Indicator Percentages*. For both events and visits, a binary matrix is constructed out of the relevant columns of each dataset. The matrix is summed row-wise into a single column vector. Each element of this vector is divided by the maximum value of the

[1]Conversions were performed using the standard tables from https://blog.prepscholar.com/act-to-sat-conversion

entire vector to obtain the *Relative Indicator Percentage* for a given student. For example, if a school offered three types of campus visits and a given student (a) attended two but there exists a student (b) who attended all three, student (a) would have a *Relative Indicator Percentage* of 67% for visits and student (b) would have a *Relative Indicator Percentage* of 100% for visits.

Some demographic variables are named and recorded the same way, but the values they take on may be meaningless in the context of another dataset. For example, a school in Pittsburgh, Pennsylvania may enroll many students whose recorded city is Pittsburgh, Pennsylvania. This does not indicate that being from Pittsburgh is indicative of enrollment in general, but rather being from a city close or far from the city where the school is located plays a role in enrollment decisions. A *Similarity Percentage* score is calculated from city, state, country and high school name columns by counting the number of admitted students in the dataset who share the same categorical value and dividing by the total number of admitted students. In the Pittsburgh example, students from Pittsburgh would have a high *Similarity Percentage* since we expect a plurality of students to be admitted from there while inversely we would expect students from Australia to have a low *Similarity Percentage*.

Some schools allowed student's to indicate which major they were interested in studying. This is another such case of where the column is consistent across schools but the values are different. These values were recorded in one of two ways: either a single variable with the name of the major or as a binary matrix of indicator variables. When it exists, the binary matrix is trivially transformed into the categorical column. Then, the *Similarity Percentage* can be calculated and is used as the transformation.

Since different customers have different ways of tracking and recording financial aid, it is necessary to consolidate variables that fall into this category in a way that can be generalized. The simplest generalization based on the provided variables is separating financial aid into institutional aid and non institutional aid. This is done by simply adding the amounts of aid explicitly noted to be institutional aid into one column and adding all other aid (ex. government aid, scholarships) into another.

2.3 Model Evaluation

2.3.1 Trade Off Between Sensitivity and Specificity

When comparing variable sets, it is necessary to determine the appropriate metric to compare instances of such sets to one another. Due to the nature of college enrollment, the large majority of accepted students will not enroll, which makes enrolling a rare event. Thus, any metric used should be able to outperform a naive classifier that predicts that every admitted student will not enroll.

In the context of college admissions, there is no intuition suggesting that a model should value identifying a true positive (a student who is predicted to enroll who actually does enroll) more than it values avoiding a false positive (a student who is predicted to enroll who does not enroll). In the former case, failing to identify the enrolling student could impose a space constraints for the college which would require the college to spend more money on housing or faculty and in the latter case the college would expect revenue it doesn't actually receive.

2.3.2 Evaluation Metric

With this in mind, we choose the Area Under the Curve of the Receiver Operating Characteristic (AUCROC) to compare models using a baseline of .5 to ensure that the model would not be beat by a naive classifier[2]. This choice of metric allows for the flexibility of the trade-off between sensitivity and specificity and is appropriate even with the sparse number of positive response variables (students enrolled)[3]. Specifically, the average AUCROC across all schools will be compared across variable sets. This will allow for the effective penalization of variables that are not shared between all of the customers because adding that variable to the Minimum Variable Set will not improve the AUCROC for a school if that school does not have that variable.

[2] An AUCROC of .5 corresponds to the 0-1 line in the ROC curve and represents a Bernoulli classifier.

[3] This is because randomly guessing via a Bernoulli distribution would result with the same probability of correctly predicting that an enrolled student will enroll as incorrectly predicting that a not enrolled student will enroll.

2.4 Variable Selection

Three selection algorithms are initially considered: a brute-force method, a forward selection method and a stochastic method. For each method, a Logistic Regression classifier is used to predict enrollment status. Logistic regression is chosen for its interpretable coefficients, however future work could easily adapt these algorithms to use other classification models.

2.4.1 Brute Force Selection of Minimum Variable Set

A brute force method can easily be seen as an infeasible solution. Consider a variable selection problem with n potential covariates. For each covariate, we have two choices: to either include or not include the variable into the final model. Thus, there are 2^n possible models that must be considered. For a variable selection problem with $n = 30$ (the approximate number of variables remaining after data cleaning and consolidation), this means there are $2^{30} \approx 10^9$ models to be considered which is unable to achieve acceptable performance in reasonable time. Thus, this method is discarded.

2.4.2 Forward Selection of Minimum Variable Set

A forward selection heuristic improves on the brute force method in that it greatly reduces the total number of models that are considered, even in the worst case. The algorithm is as follows:

In each step of the algorithm, a random variable not in the Minimum Variable Set is considered. If adding that variable to the minimum variable set improves the average AUCROC across all datasets, that variable is added to the minimum variable set. If not, the variable is discarded. The process completes until all variables have been considered. The main limitation of this algorithm is that it is greedy and can possibly underperform based on the random order on which variables are added into the model.

Algorithm 1 Forward Model Selection

Let $\overline{C}$ be the set of all available variables
Let score be a function of a set that returns the average AUCROC across all datasets D
$C \leftarrow$ randomize($\overline{C}$)
Initialize $C^* = \emptyset$ be the MVS
Initialize $m^* = .5$ be the best achieved average AUCROC for the MVS
for all $c \in C$ **do**
  $m \leftarrow$ score$(C^* \cup \{c\}, D)$
  if $m > m^*$ **then**
    $m^* \leftarrow \overline{m}$
    $C^* \leftarrow C^* \cup \{c\}$
  end if
end for

2.4.3 Stochastic Selection of Minimum Variable Set

We propose a stochastic variable selection method to narrow down the variables remaining after variable removal and data cleaning using an evolutionary algorithm. This process allows for a non greedy selection of the minimum variable set for a given customer and model by considering many orders of magnitude fewer models than a brute force method.

Algorithm 2 Stochastic Model Selection

Let C be the set of all available variables
Let score be a function of a set that returns the average AUCROC across all datasets D
Initialize $s^* = .5$ be the best achieved average AUCROC for the MVS
Let G be the total number of evolution generations
Let N be the total number of models considered in each generation
Let $M = \emptyset$ be the models "alive" in the current generation

Let f be the percentage of fit models that survive each generation
Let u be the percentage of randomly selected unfit models that survice each generation
Let z be the probability of mutation

Let random_subset be a function that takes in a set and returns a random subset
Let take_fit_models be a function that takes in a sorted set S and a percentage f and returns the first $f\%$ of elements of S
Let take_fit_models be a function that takes in a sorted set S and a percentage u and returns a random $u\%$ of elements of S
Let create_parent_models be a function that takes in a set S and returns pairings of elements of S in two sets S_1 and S_2
Let create_random_combination be a function that takes two models m_1, m_2 and returns a new model m that has variable inclusion settings that are randomly chosen from the variable inclusion settings of m_1 and m_2
Let should_mutate be a function that takes in a probability p and returns True if a random number generated is less than p
Let mutate be a function that takes in a model m and randomly changes a variable inclusion setting.

for $n \in [1, N]$ **do**
  $\overline{M} \leftarrow$ random_subset(C)
  $M \leftarrow M \cup \{\overline{M}\}$
end for
for $g \in [1, g]$ **do**
  $S = \emptyset$
  for all $m \in M$ **do**
    $s \leftarrow$ score(m, D)
    $S \leftarrow S \cup \{s\}$
  end for
  $\overline{s} \leftarrow$ mean(S)
  if $\overline{s} \leq s^*$ **then**
    break
  end if
  $M \leftarrow$ sorted(M, S)
  $M^* = \emptyset$
  $M^* \leftarrow M^* \cup$ take_fit_models(M, f)
  $M^* \leftarrow M^* \cup$ take_unfit_models(M, u)
  $P_1, P_2 \leftarrow$ create_parent_models(M^*)
  for all $p_1, p_2 \in P_1, P_2$ **do**
    if length$(M^*) < N$ **then**
      $p^* \leftarrow$ create_random_combination(p_1, p_2)
      if should_mutate(z) **then**
        $p^* \leftarrow$ mutate(p^*)
      end if
      $M^* \leftarrow M^* \cup \{p^*\}$
    end if
  end for
  $M \leftarrow M^*$
end for

For a given school, we randomize a population of unique variable sets based on the variables available for that school. In each iteration, a given model is fit to each set in the population and 3-fold cross validation of the AUCROC metric is scored against every dataset and then averaged. The best performing models and a small random selection of models that did not perform as well survive. Some of these models (parents) are randomly selected to breed, which creates new unique sets of variables (children) whose elements are determined by the elements of the parents. Also, some of these sets undergo a mutation, where a random variable is chosen to either be included or excluded despite whether that variables status in the parent sets. These sets are combined with the surviving sets from the previous iteration and the process repeats with these sets until the average AUCROC metric across all models fails to improve from the previous generation.

2.5 Factor Analysis and Principal Components Analysis

2.5.1 Factor Analysis

Factor Analysis with four factors is used to determine whether the hypothesized categories are actually latent variables that explain college enrollment for a given student. Plots of factor loadings are used to determine visually whether the four factors are contributing to the same variables in a given dataset regardless of whether the full variable set or the Minimum Variable Set is used. In addition, these factors are compared across schools to validate whether or not there is a generalizable relationship.

2.5.2 Principal Components Analysis

Principal Components Analysis with four components is used to validate that there is not a substantial dropoff in the proportion of variance in the data explained by four components when only using the Minimum Variable Set as a subset of the data.

3 Results

3.1 Minimum Variable Set

3.1.1 Stepwise Variable Selection

The Minimum Variable Set as selected by the stepwise (greedy) algorithm is:

- Visit
- Family Type Max
- Education Max
- Event
- State
- Interested in Campus Housing
- Average Household Size
- High School Percentile
- Earnings Max
- Distance
- City
- Initial Source of Contact
- US Citizenship Status
- Sex
- Max Test Score
- Legacy Status
- High School GPA

- Interested in Receiving Financial Aid
- Median Age

The stepwise algorithm achieved an average AUCROC of 75.46% across all schools and found a set of 21 variables.

3.1.2 Stochastic Variable Selection

The Minimum Variable Set as selected by the stochastic algorithm is:

- Institutional Aid
- Initial Source of Contact
- Visit
- Median Age
- Event
- Family Type Max
- Max Test Score
- City
- High School GPA
- High School Class Ranking
- Sex
- Interested in Campus Housing
- Distance
- Mean Usual Hours
- Average Household Size
- State
- Major

Academics	Demographics	Interest	Financial Aid
High School Class Rank	Sex	Major	Institutional Aid Amount
Max Test Score	Distance	Visits	
High School GPA	City	Wants Campus Housing	
	Mean Usual Hours	Events	
	Family Type Max	Initial Source of Contact	
	Average Household Size		
	Median Age		
	State		

Table 4: Breakdown of Minimum Variable Set into four proposed categories

The stochastic algorithm achieved an average AUCROC of 76% across all schools and found a Minimum Variable Set of 17 variables.

3.1.3 ROC Curve

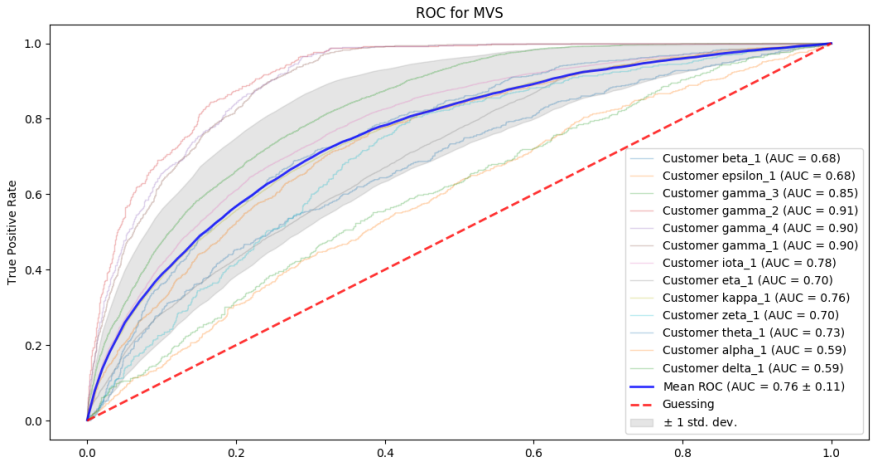


Figure 7: ROC Curves for the Minimum Variable Set chosen by the Stochastic Algorithm

The ROC Curves in Figure 7 demonstrate that the Minimum Variable Set is able to outperform guessing in terms of the AUCROC metric.

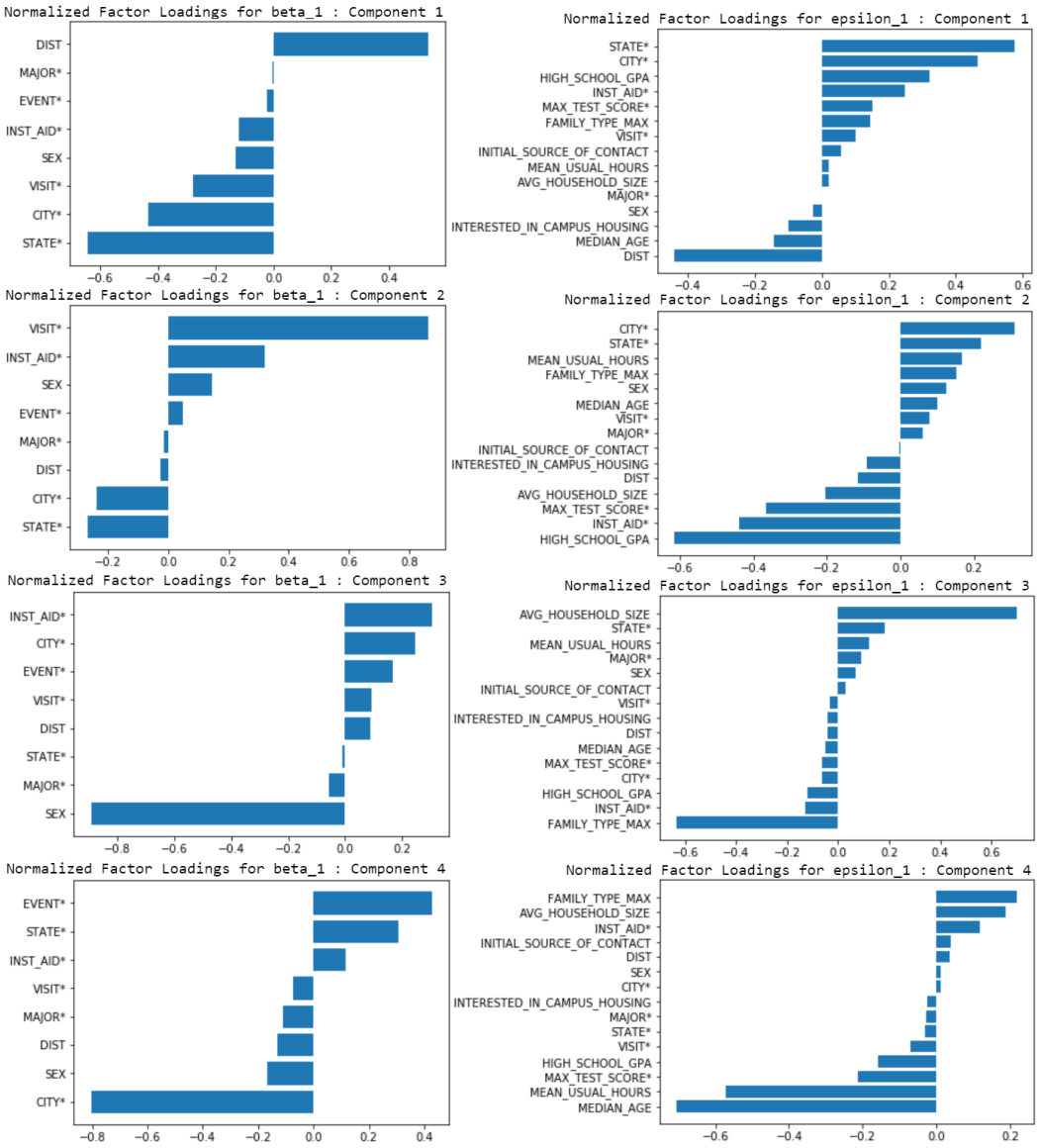
Customer	AUCROC
Alpha 1	59%
Beta 1	68%
Gamma 1	**90%**
Gamma 2	**91%**
Gamma 3	**85%**
Gamma 4	**90%**
Delta 1	59%
Epsilon 1	68%
Zeta 1	70%
Eta 1	70%
Theta 1	73%
Iota 1	78%
Kappa 1	76%

Table 5: MVS AUCROC Performance by Customer

3.2 Diagnostics

3.2.1 Factor Analysis

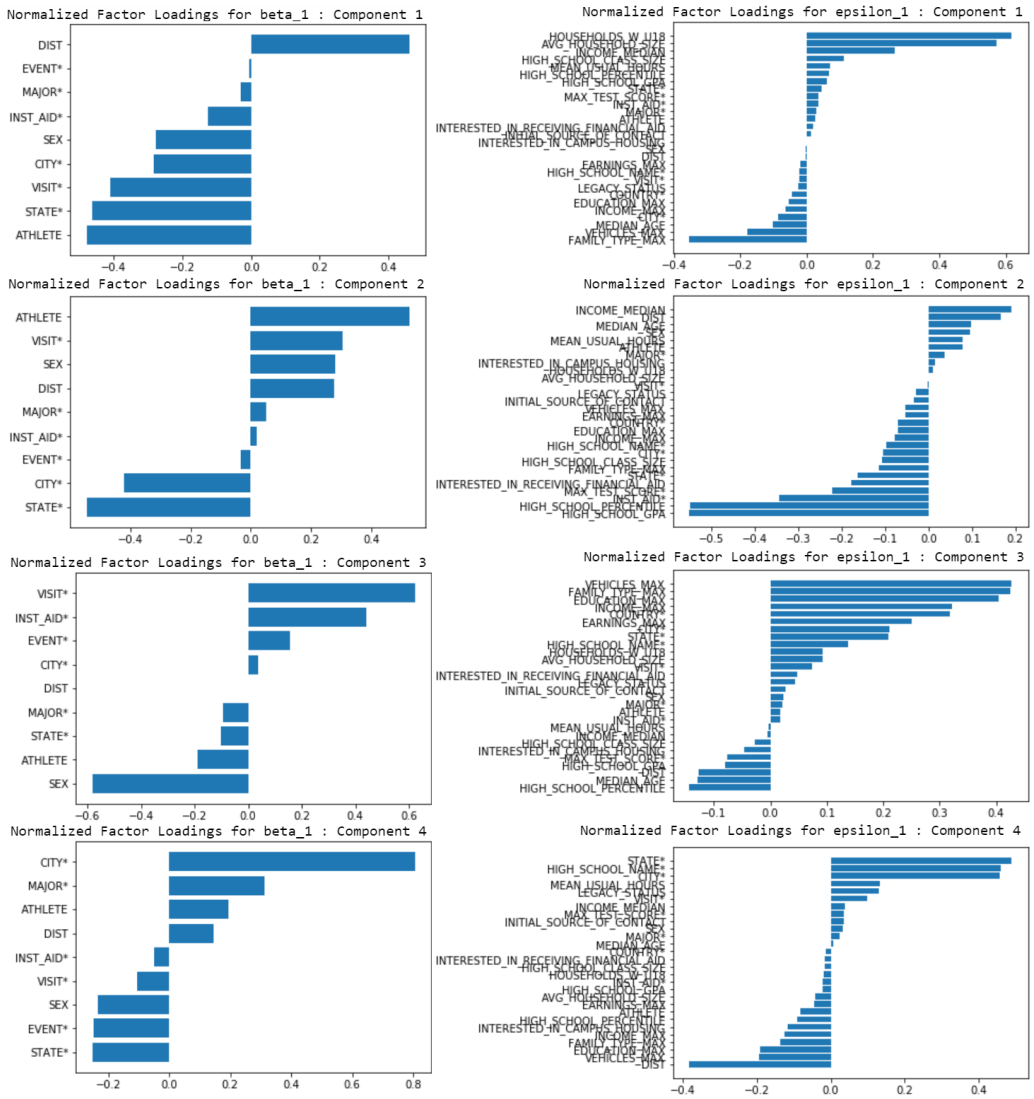
The results of the factor analysis on the Minimum Variable Set data are shown in Figure 8 compare the normalized factor loadings between customers, using Beta 1 and Epsilon 1 as an example. Note that due to the drastic difference in the variables available in each dataset, its hard to establish the four proposed categories as shared latent variables between the two schools.



Normalized Factor Loadings for Beta 1 and Epsilon 1 using only variables from minimum variable set

Figure 8: Factor Analysis with MVS

However, when comparing the previous factor loadings to the factor loadings from the same school using the full variable set, there are some patterns that visually emerge. Plots of normalized factor loadings for the full variable set for Beta 1 and Epsilon 1 are displayed in Figure 9 For example, in Beta 1, the component 1 in each set look fairly similar, with Distance on top and City, State and Visits on bottom.



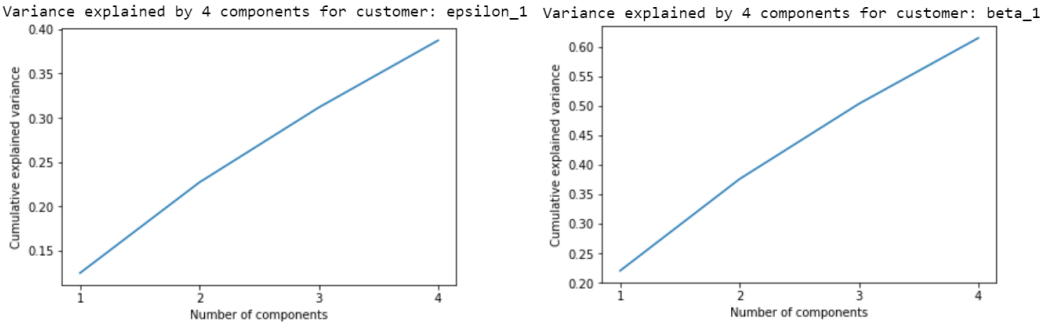
Normalized Factor Loadings for Beta 1 and Epsilon 1 using all variables in admitted life cycle

Figure 9: Factor Analysis with full variable set

While some similarities in the loadings exist, there is not strong enough evidence to definitively conclude that the latent variables of academics, demographics, interests and financial aid are the sole latent variables that are driving enrollment decisions. Thus, while its possible to establish a Minimum Variable Set that appears to capture the same underlying factors within a given school, these factors vary from school to school and are not identifiable.

3.2.2 Principal Components Analysis

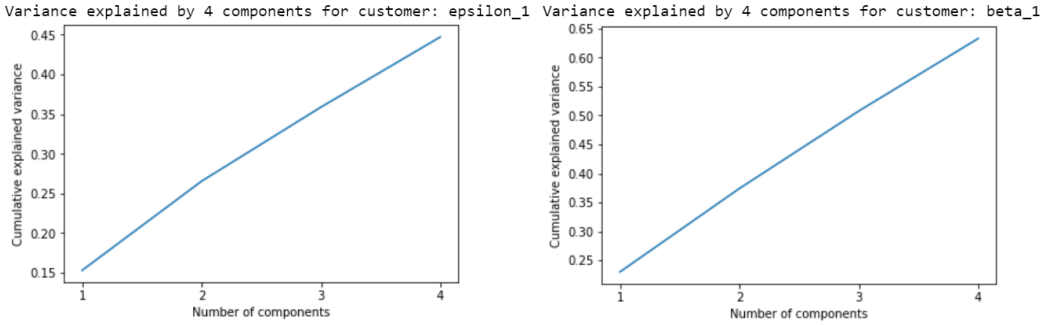
Figure 10 shows the proportion of variance explained by four principal components, using the full variable set. Notice that for Epsilon 1, those four components explain a little less than 40% of the variability in the data while for Beta 1, they explain about 63%.



Variance explained by 4 components (from all variables in admitted lifecycle)

Figure 10: Variance explained with four components - full variable set

Compare these plots to the plots of the proportion of variance explained by four components when we restrict the analysis to the Minimum Variable Set in Figure 11. The first four components of Epsilon 1 explain about 45% of the variability of the data while the first four components for Beta 1 explain 65% of the variability. Thus, we see that these components are slightly better in capturing the variability of the data than when we use all of the variables in the given admitted lifecycle.



Variance explained by 4 components (from variables in stochastic model only)

Figure 11: Proportion of variance explained by four components - MVS

These findings validate the use of the Minimum Variable Set in that there is not a dropoff in the proportion of variability explained after removing most variables in the first four principal components. This means that the variables included in the set are not linear combinations of one another and represent different information.

4 Discussion

4.1 Analysis

Ultimately, both the stepwise algorithm and the stochastic algorithm are able to identify Minimum Variable Sets that achieve similar performance using average AUCROC across all schools as a comparison metric. The stochastic selection process is able to achieve a better average AUCROC with less variables and thus the set generated by that process is considered to be the Minimum Variable Set. Table 5 shows how the Minimum Variable Set is able to successfully cover each of the four main categories we decided were necessary to cover for a Minimum Variable Set to be considered valid.

In the academics category, the included variables are High School Class Rank, Max Test Score and High School GPA. These variables paint a fairly complete picture of a student's academic ability, as standardized test scores can't be inflated by a school, GPA serves to benefit those students who don't test well and ranking gives an idea of where a student stands relative to peers in their own school.

In the demographics category, there are a few distinct groups that form (which may explain why a factor model with four factors was insufficient). Distance, City and State all relate to how far a student is from a given school. These will obviously play a role in an enrollment decision because students will most likely attend schools that are closer and that people who they know have also attended. The other demographic variables, Mean Usual Hours, Family Type Max, Average Household Size and Median Age all speak to census level information that can serve as proxies for family income and expectations regarding higher education. It's worth noting that in many of the top performing variable sets that the census variables appear interchangeably. This suggests that while it is necessary to include some of these variables to obtain proxies for the properties of the community where a student is from, none of these types of variables specifically are more important than the others. The last variable, Sex, is difficult to understand why it would end up in the Minimum Variable Set. One explanation may be that certain programs offered by the schools have more appeal to one gender compared to the other which results in higher enrollment rates school-wide for that gender.

Like academics, the interest level variables that made it into the Minimum Variable Set are fairly easy to explain. Attending campus Visits and Events is indicative of a desire to attend that school because attendance costs both time and money. Whether or not a student wants to live on campus may have different effects depending on the school. At some schools, students may value being able to live at home while at others wanting to live on campus may suggest a desire to integrate with the university culture and other students who attend. Either way, it makes sense that this variable would be important across schools. The Initial Source of Contact that a school has with a student also seems to be reasonably informative. Students who make first contact on their own by visiting or going to an event are probably more likely to enroll than a student who just applied and never made any effort to connect with the school.

What is most interesting about the financial aid category, is that despite the expectation that it would be the most important category for predicting enrollment and yet only Institutional Aid was included in the Minimum Variable Set. This may partly be because many of the financial aid level variables were consolidated as a result of the data cleaning process. One reason why Institutional Aid may have been included is that Government Aid would be consistent across all schools that a given student applied to. This means that the student is really only considering the cost of the school minus whatever government money they are receiving and the financial aid that the school itself offers them.

The mean ROC Curve lies above the guessing line which provides strong evidence that the Minimum Variable Set is able to uncover structure within the data to effectively separate students who will enroll and students who will not enroll.

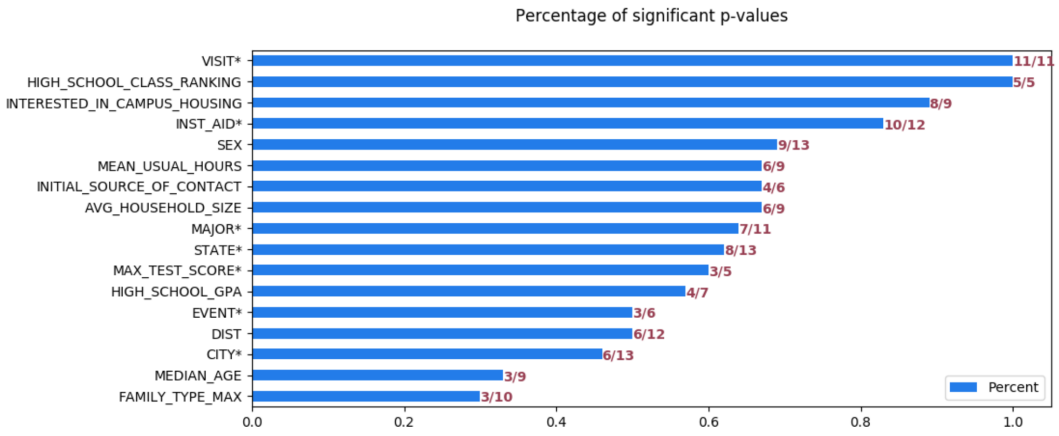


Figure 12: Proportion of times that a variable in the MVS was significant at the $\alpha = .05$ level for logistic regressions fit to each dataset

Logistic regression models are fit to each dataset using only the available variables that are in the Minimum Variable Set. Of the variables in the Minimum Variable Set, while all were statistically

significant at the $\alpha = .05$ level for at least one customer, not all were statistically significant in every dataset that they appeared in. Figure 12 exemplifies this by plotting the proportion of times a variable was statistically significant (where the denominator is the number of schools where the variable was available). Visits and Institutional Aid stand out as both appearing in most datasets and being significant in a large proportion (if not all) datasets that they appear in. This indicates that visits and institutional aid may be some of the most important variables for predicting enrollment. This can be explained by the fact that students who visit a school (which costs time and money) have demonstrated interest in attending the school and if they receive enough aid they can afford to attend. Variables such as Median Age and Family Type Max don't typically appear as significant even though they appear in many datasets. Most likely this is because these variables are not predictive on their own and this suggests that these variables should be swapped out for census level variables that might do a better job at predicting enrollment.

4.2 Limitations

One caveat to these findings is that since customer Gamma has four campuses that are considered as separate schools, these sets will be biased towards the factors that influence enrollment at Gamma campuses. Notice that, in the plot of ROC in Figure 7, each Gamma campus achieved a cross validated AUCROC at or greater than one standard deviation above the AUC for the mean ROC curve. This suggests that this bias is present in the data and would need to be corrected for in some way in future work.

In addition, there were no explored interactions between the variables despite the fact that it's very reasonable for some factors to vary based on others. This could be responsible for excluding variables from the Minimum Variable Set that actually would have better performance than other included variables.

4.3 Future Work

While this analysis was able to successfully identify a Minimum Variable Set, there are a variety of avenues that were unable to be explored in depth that could still be of interest to better understand the factors that influence student enrollment decisions.

For example, this investigation focuses on the information available only in the admitted lifecycles for each school. However, it is reasonable to assume that different lifecycles would have their own appropriate Minimum Variable Set, since these lifecycles contain different levels of information as well as different student populations. An analogous analysis could be performed on other lifecycles that these schools believe would provide better insight into what is associated with student enrollment.

Also, as a result of requiring a variable to be included in four schools before it could be considered for the Minimum Variable Set, variables that were specific to a certain school were naturally excluded prior to the variable selection algorithm. Furthermore, by averaging the AUCROC metric across each school, the selection algorithms have a bias toward choosing variables that are common to most schools. These less common variables were excluded because they didn't generalize to the datasets as a whole but they may be extremely useful for predicting enrollment for the schools that do have them. Future iterations of Minimum Variable Set selection should try to identify less common variables that are very predictive for the schools that contain them and add them to the general set. This could greatly improve model performance while still keeping the total number of variables in the set relatively small.

Even among the variables that met that threshold criteria, this analysis does not deeply explore the differences in the granularity of all the variables. For example, further exploration could have revealed that it is not necessary to have multiple visit indicators and the real difference is between whether the visit was on-campus or off-campus.

5 Appendix

5.1 Minimum Variable Set Model Summaries

Alpha 1 - Logistic Regression

Dep. Variable:	ENROLLMENT_STATUS	**No. Observations:**	3342
Model:	Logit	**Df Residuals:**	3329
Method:	MLE	**Df Model:**	12
Date:	Wed, 02 May 2018	**Pseudo R-squ.:**	0.1358
Time:	20:24:06	**Log-Likelihood:**	-1451.9
converged:	True	**LL-Null:**	-1680.1

	Coef	**Std Err**	**z**	**P>|z|**	**[0.025**	**0.975]**
AVG_HOUSEHOLD_SIZE	-0.4154	0.125	-3.332	0.001	-0.660	-0.171
DIST	-0.0002	0.000	-1.512	0.131	-0.000	5.26e-05
FAMILY_TYPE_MAX	-0.1709	0.056	-3.076	0.002	-0.280	-0.062
INITIAL_SOURCE_OF_CONTACT	0.0170	0.006	3.016	0.003	0.006	0.028
MEAN_USUAL_HOURS	0.0084	0.017	0.505	0.613	-0.024	0.041
MEDIAN_AGE	-0.0096	0.009	-1.022	0.307	-0.028	0.009
SEX	0.1637	0.094	1.749	0.080	-0.020	0.347
INST_AID*	-3.388e-05	9.48e-06	-3.574	0.000	-5.25e-05	-1.53e-05
VISIT*	4.6784	0.286	16.382	0.000	4.119	5.238
MAX_TEST_SCORE*	-0.0001	0.000	-0.425	0.671	-0.001	0.000
CITY*	-5.1987	2.069	-2.513	0.012	-9.253	-1.144
STATE*	0.3094	0.634	0.488	0.626	-0.933	1.552
MAJOR*	-3.1919	1.515	-2.108	0.035	-6.160	-0.223

Beta 1 - Logistic Regression

Dep. Variable:	ENROLLMENT_STATUS	**No. Observations:**	2247
Model:	Logit	**Df Residuals:**	2239
Method:	MLE	**Df Model:**	7
Date:	Wed, 02 May 2018	**Pseudo R-squ.:**	0.07941
Time:	20:24:06	**Log-Likelihood:**	-1221.2
converged:	True	**LL-Null:**	-1326.5

	Coef	**Std Err**	**z**	**P>|z|**	**[0.025**	**0.975]**
DIST	-0.0033	0.000	-11.037	0.000	-0.004	-0.003
SEX	-0.3257	0.097	-3.371	0.001	-0.515	-0.136
INST_AID*	3.259e-06	4.91e-06	0.663	0.507	-6.37e-06	1.29e-05
VISIT*	1.5076	0.172	8.742	0.000	1.170	1.846
EVENT*	0.8611	0.344	2.501	0.012	0.186	1.536
CITY*	8.9492	6.815	1.313	0.189	-4.408	22.307
STATE*	-0.9925	0.240	-4.142	0.000	-1.462	-0.523
MAJOR*	-3.7771	0.868	-4.352	0.000	-5.478	-2.076

Gamma 1 - Logistic Regression

Dep. Variable:	ENROLLMENT_STATUS	**No. Observations:**	7002
Model:	Logit	**Df Residuals:**	6986
Method:	MLE	**Df Model:**	15
Date:	Wed, 02 May 2018	**Pseudo R-squ.:**	0.3894
Time:	20:24:06	**Log-Likelihood:**	-1871.2
converged:	True	**LL-Null:**	-3064.7

	Coef	**Std Err**	**z**	**P>|z|**	**[0.025**	**0.975]**
AVG_HOUSEHOLD_SIZE	-0.5218	0.121	-4.329	0.000	-0.758	-0.286
DIST	0.0002	0.000	1.526	0.127	-7.02e-05	0.001
FAMILY_TYPE_MAX	-0.1236	0.033	-3.757	0.000	-0.188	-0.059
HIGH_SCHOOL_CLASS_RANKING	0.0011	0.000	2.909	0.004	0.000	0.002
HIGH_SCHOOL_GPA	0.0180	0.049	0.364	0.716	-0.079	0.115
INITIAL_SOURCE_OF_CONTACT	-0.1581	0.043	-3.693	0.000	-0.242	-0.074
INTERESTED_IN_CAMPUS_HOUSING	1.1565	0.121	9.573	0.000	0.920	1.393
MEAN_USUAL_HOURS	-0.0927	0.014	-6.855	0.000	-0.119	-0.066
MEDIAN_AGE	-0.0245	0.008	-3.268	0.001	-0.039	-0.010
SEX	-0.3041	0.088	-3.462	0.001	-0.476	-0.132
INST_AID*	0.0002	7.75e-06	25.282	0.000	0.000	0.000
VISIT*	3.5610	0.187	18.999	0.000	3.194	3.928
EVENT*	-0.6127	0.149	-4.114	0.000	-0.905	-0.321
CITY*	15.1099	3.848	3.927	0.000	7.568	22.651
STATE*	1.3067	0.284	4.605	0.000	0.751	1.863
MAJOR*	3.9599	0.497	7.975	0.000	2.987	4.933

Gamma 2 - Logistic Regression

Dep. Variable:	ENROLLMENT_STATUS	**No. Observations:**	4319
Model:	Logit	**Df Residuals:**	4303
Method:	MLE	**Df Model:**	15
Date:	Wed, 02 May 2018	**Pseudo R-squ.:**	0.4047
Time:	20:24:06	**Log-Likelihood:**	-973.53
converged:	True	**LL-Null:**	-1635.3

	Coef	**Std Err**	**z**	**P>|z|**	**[0.025**	**0.975]**
AVG_HOUSEHOLD_SIZE	-0.2753	0.117	-2.343	0.019	-0.506	-0.045
DIST	7.417e-05	0.000	0.540	0.589	-0.000	0.000
FAMILY_TYPE_MAX	-0.0852	0.054	-1.580	0.114	-0.191	0.020
HIGH_SCHOOL_CLASS_RANKING	0.0013	0.000	2.802	0.005	0.000	0.002
HIGH_SCHOOL_GPA	-0.1460	0.072	-2.029	0.042	-0.287	-0.005
INITIAL_SOURCE_OF_CONTACT	-0.1380	0.058	-2.399	0.016	-0.251	-0.025
INTERESTED_IN_CAMPUS_HOUSING	-2.3213	0.183	-12.700	0.000	-2.680	-1.963
MEAN_USUAL_HOURS	-0.0654	0.017	-3.799	0.000	-0.099	-0.032
MEDIAN_AGE	0.0121	0.011	1.067	0.286	-0.010	0.034
SEX	0.4429	0.121	3.652	0.000	0.205	0.681
INST_AID*	0.0002	8.55e-06	18.043	0.000	0.000	0.000
VISIT*	3.9216	0.281	13.972	0.000	3.371	4.472
EVENT*	0.3619	0.282	1.281	0.200	-0.192	0.916
CITY*	4.2755	4.671	0.915	0.360	-4.880	13.431
STATE*	1.0409	1.006	1.034	0.301	-0.931	3.013
MAJOR*	3.4792	0.915	3.804	0.000	1.687	5.272

Gamma 3 - Logistic Regression

Dep. Variable:	ENROLLMENT_STATUS	**No. Observations:**	23367
Model:	Logit	**Df Residuals:**	23351
Method:	MLE	**Df Model:**	15
Date:	Wed, 02 May 2018	**Pseudo R-squ.:**	0.2706
Time:	20:24:06	**Log-Likelihood:**	-7279.1
converged:	True	**LL-Null:**	-9979.1

	Coef	**Std Err**	**z**	**P>|z|**	**[0.025**	**0.975]**
AVG_HOUSEHOLD_SIZE	-0.4267	0.057	-7.525	0.000	-0.538	-0.316
DIST	-0.0001	4.88e-05	-2.702	0.007	-0.000	-3.62e-05
FAMILY_TYPE_MAX	-0.0330	0.018	-1.837	0.066	-0.068	0.002
HIGH_SCHOOL_CLASS_RANKING	0.0019	0.000	8.543	0.000	0.001	0.002
HIGH_SCHOOL_GPA	-0.0474	0.025	-1.893	0.058	-0.096	0.002
INITIAL_SOURCE_OF_CONTACT	-0.1537	0.017	-9.212	0.000	-0.186	-0.121
INTERESTED_IN_CAMPUS_HOUSING	1.8080	0.067	27.106	0.000	1.677	1.939
MEAN_USUAL_HOURS	-0.0903	0.007	-12.661	0.000	-0.104	-0.076
MEDIAN_AGE	0.0064	0.004	1.490	0.136	-0.002	0.015
SEX	-0.4191	0.043	-9.756	0.000	-0.503	-0.335
INST_AID*	0.0001	3.41e-06	34.424	0.000	0.000	0.000
VISIT*	4.8601	0.120	40.540	0.000	4.625	5.095
EVENT*	0.0190	0.105	0.180	0.857	-0.187	0.225
CITY*	-5.8994	4.040	-1.460	0.144	-13.817	2.018
STATE*	0.8555	0.458	1.870	0.062	-0.041	1.752
MAJOR*	3.3588	0.712	4.717	0.000	1.963	4.755

Gamma 4 - Logistic Regression

Dep. Variable:	ENROLLMENT_STATUS	**No. Observations:**	6636
Model:	Logit	**Df Residuals:**	6620
Method:	MLE	**Df Model:**	15
Date:	Wed, 02 May 2018	**Pseudo R-squ.:**	0.3764
Time:	20:24:06	**Log-Likelihood:**	-1488.0
converged:	True	**LL-Null:**	-2386.0

	Coef	**Std Err**	**z**	**P>|z|**	**[0.025**	**0.975]**
AVG_HOUSEHOLD_SIZE	-0.1191	0.105	-1.137	0.256	-0.325	0.086
DIST	-0.0003	0.000	-2.048	0.041	-0.001	-1.33e-05
FAMILY_TYPE_MAX	0.0201	0.034	0.586	0.558	-0.047	0.087
HIGH_SCHOOL_CLASS_RANKING	0.0013	0.000	3.139	0.002	0.000	0.002
HIGH_SCHOOL_GPA	-0.1315	0.054	-2.433	0.015	-0.237	-0.026
INITIAL_SOURCE_OF_CONTACT	-0.0804	0.051	-1.573	0.116	-0.181	0.020
INTERESTED_IN_CAMPUS_HOUSING	-1.2638	0.133	-9.479	0.000	-1.525	-1.002
MEAN_USUAL_HOURS	-0.0944	0.014	-6.755	0.000	-0.122	-0.067
MEDIAN_AGE	-0.0037	0.008	-0.439	0.661	-0.020	0.013
SEX	0.1081	0.092	1.179	0.238	-0.072	0.288
INST_AID*	0.0002	9.01e-06	23.505	0.000	0.000	0.000
VISIT*	4.3018	0.279	15.397	0.000	3.754	4.849
EVENT*	-0.1975	0.185	-1.066	0.286	-0.561	0.166
CITY*	-14.6534	5.237	-2.798	0.005	-24.917	-4.390
STATE*	0.8432	0.478	1.765	0.077	-0.093	1.779
MAJOR*	0.3892	0.653	0.596	0.551	-0.891	1.670

Delta 1 - Logistic Regression

Dep. Variable:	ENROLLMENT_STATUS	**No. Observations:**	1930
Model:	Logit	**Df Residuals:**	1917
Method:	MLE	**Df Model:**	12
Date:	Wed, 02 May 2018	**Pseudo R-squ.:**	0.1070
Time:	20:24:06	**Log-Likelihood:**	-1039.6
converged:	True	**LL-Null:**	-1164.2

	Coef	**Std Err**	**z**	**P>|z|**	**[0.025**	**0.975]**
AVG_HOUSEHOLD_SIZE	0.0029	0.194	0.015	0.988	-0.378	0.384
DIST	-0.0003	0.000	-0.949	0.343	-0.001	0.000
FAMILY_TYPE_MAX	-0.0747	0.065	-1.148	0.251	-0.202	0.053
INTERESTED_IN_CAMPUS_HOUSING	-0.2815	0.161	-1.747	0.081	-0.597	0.034
MEAN_USUAL_HOURS	-0.0012	0.020	-0.061	0.951	-0.041	0.039
MEDIAN_AGE	0.0026	0.011	0.235	0.815	-0.019	0.024
SEX	0.2625	0.111	2.359	0.018	0.044	0.481
INST_AID*	-4.669e-05	2.03e-05	-2.303	0.021	-8.64e-05	-6.95e-06
VISIT*	4.3328	0.330	13.126	0.000	3.686	4.980
MAX_TEST_SCORE*	-0.0008	0.000	-1.663	0.096	-0.002	0.000
CITY*	-0.9889	0.613	-1.612	0.107	-2.191	0.214
STATE*	0.0996	0.251	0.397	0.692	-0.392	0.592
MAJOR*	-1.4939	0.905	-1.651	0.099	-3.268	0.280

Epsilon 1 - Logistic Regression

Dep. Variable:	ENROLLMENT_STATUS	**No. Observations:**	5546
Model:	Logit	**Df Residuals:**	5531
Method:	MLE	**Df Model:**	14
Date:	Wed, 02 May 2018	**Pseudo R-squ.:**	0.1361
Time:	20:24:07	**Log-Likelihood:**	-2342.3
converged:	True	**LL-Null:**	-2711.4

	Coef	**Std Err**	**z**	**P>|z|**	**[0.025**	**0.975]**
AVG_HOUSEHOLD_SIZE	0.1669	0.103	1.622	0.105	-0.035	0.368
DIST	8.486e-06	0.000	0.080	0.936	-0.000	0.000
FAMILY_TYPE_MAX	0.0939	0.049	1.932	0.053	-0.001	0.189
HIGH_SCHOOL_GPA	-0.6734	0.116	-5.797	0.000	-0.901	-0.446
INITIAL_SOURCE_OF_CONTACT	0.0049	0.004	1.174	0.240	-0.003	0.013
INTERESTED_IN_CAMPUS_HOUSING	-1.2550	0.189	-6.642	0.000	-1.625	-0.885
MEAN_USUAL_HOURS	0.0186	0.013	1.403	0.161	-0.007	0.044
MEDIAN_AGE	0.0064	0.006	0.995	0.320	-0.006	0.019
SEX	0.0901	0.077	1.168	0.243	-0.061	0.241
INST_AID*	6.4e-05	4.38e-06	14.604	0.000	5.54e-05	7.26e-05
VISIT*	1.4871	0.082	18.054	0.000	1.326	1.649
MAX_TEST_SCORE*	-0.0015	0.000	-3.967	0.000	-0.002	-0.001
CITY*	-2.3634	3.538	-0.668	0.504	-9.298	4.571
STATE*	2.8849	0.616	4.680	0.000	1.677	4.093
MAJOR*	1.1177	0.760	1.470	0.141	-0.372	2.608

Zeta 1 - Logistic Regression

Dep. Variable:	ENROLLMENT_STATUS	**No. Observations:**	2053
Model:	Logit	**Df Residuals:**	2045
Method:	MLE	**Df Model:**	7
Date:	Wed, 02 May 2018	**Pseudo R-squ.:**	0.1312
Time:	20:24:07	**Log-Likelihood:**	-1134.5
converged:	True	**LL-Null:**	-1305.8

	Coef	**Std Err**	**z**	**P>|z|**	**[0.025**	**0.975]**
DIST	-0.0008	0.000	-4.899	0.000	-0.001	-0.000
FAMILY_TYPE_MAX	-0.0432	0.037	-1.179	0.239	-0.115	0.029
INTERESTED_IN_CAMPUS_HOUSING	-0.8522	0.117	-7.282	0.000	-1.082	-0.623
SEX	0.1692	0.114	1.480	0.139	-0.055	0.393
INST_AID*	-2.038e-05	1.39e-05	-1.466	0.143	-4.76e-05	6.87e-06
CITY*	1.3551	15.676	0.086	0.931	-29.370	32.080
STATE*	2.5798	0.298	8.661	0.000	1.996	3.164
MAJOR*	-0.5779	0.134	-4.321	0.000	-0.840	-0.316

Eta 1 - Logistic Regression

Dep. Variable:	ENROLLMENT_STATUS	**No. Observations:**	36031
Model:	Logit	**Df Residuals:**	36022
Method:	MLE	**Df Model:**	8
Date:	Wed, 02 May 2018	**Pseudo R-squ.:**	0.09857
Time:	20:24:07	**Log-Likelihood:**	-22143.
converged:	True	**LL-Null:**	-24564.

	Coef	**Std Err**	**z**	**P>|z|**	**[0.025**	**0.975]**
AVG_HOUSEHOLD_SIZE	-0.2105	0.024	-8.752	0.000	-0.258	-0.163
FAMILY_TYPE_MAX	-0.0123	0.013	-0.965	0.335	-0.037	0.013
MEAN_USUAL_HOURS	-0.0316	0.004	-8.624	0.000	-0.039	-0.024
MEDIAN_AGE	-0.0145	0.002	-7.350	0.000	-0.018	-0.011
SEX	-0.2775	0.023	-12.092	0.000	-0.322	-0.233
EVENT*	5.0499	0.137	36.849	0.000	4.781	5.318
CITY*	1.0436	0.202	5.164	0.000	0.648	1.440
STATE*	3.2517	0.073	44.717	0.000	3.109	3.394
MAJOR*	0.0886	0.937	0.095	0.925	-1.748	1.926

Theta 1 - Logistic Regression

Dep. Variable:	ENROLLMENT_STATUS	**No. Observations:**	3123
Model:	Logit	**Df Residuals:**	3114
Method:	MLE	**Df Model:**	8
Date:	Wed, 02 May 2018	**Pseudo R-squ.:**	0.2320
Time:	20:24:07	**Log-Likelihood:**	-1404.0
converged:	True	**LL-Null:**	-1828.2

	Coef	**Std Err**	**z**	**P>|z|**	**[0.025**	**0.975]**
DIST	-0.0004	0.000	-1.807	0.071	-0.001	3.09e-05
HIGH_SCHOOL_GPA	-0.0019	0.001	-1.377	0.169	-0.005	0.001
INTERESTED_IN_CAMPUS_HOUSING	-1.5100	0.170	-8.861	0.000	-1.844	-1.176
MAX_TEST_SCORE*	-0.0048	0.000	-15.421	0.000	-0.005	-0.004
SEX	-0.3470	0.096	-3.607	0.000	-0.536	-0.158
INST_AID*	0.0002	1.39e-05	15.866	0.000	0.000	0.000
VISIT*	5.2409	0.298	17.577	0.000	4.657	5.825
CITY*	-1.9451	2.318	-0.839	0.401	-6.489	2.599
STATE*	0.7848	0.217	3.622	0.000	0.360	1.210

Iota 1 - Logistic Regression

Dep. Variable:	ENROLLMENT_STATUS	**No. Observations:**	24980
Model:	Logit	**Df Residuals:**	24969
Method:	MLE	**Df Model:**	10
Date:	Wed, 02 May 2018	**Pseudo R-squ.:**	0.1763
Time:	20:24:07	**Log-Likelihood:**	-13052.
converged:	True	**LL-Null:**	-15846.

	Coef	**Std Err**	**z**	**P>|z|**	**[0.025**	**0.975]**
AVG_HOUSEHOLD_SIZE	-0.3099	0.039	-7.848	0.000	-0.387	-0.233
DIST	-0.0015	0.000	-14.429	0.000	-0.002	-0.001
FAMILY_TYPE_MAX	-0.0774	0.025	-3.136	0.002	-0.126	-0.029
HIGH_SCHOOL_CLASS_RANKING	0.0022	0.000	21.616	0.000	0.002	0.002
MEAN_USUAL_HOURS	0.0191	0.005	4.187	0.000	0.010	0.028
MEDIAN_AGE	-0.0138	0.003	-4.775	0.000	-0.020	-0.008
SEX	-0.1036	0.030	-3.480	0.001	-0.162	-0.045
INST_AID*	0.0004	1.49e-05	30.109	0.000	0.000	0.000
VISIT*	5.4767	0.113	48.468	0.000	5.255	5.698
CITY*	-6.5498	0.730	-8.975	0.000	-7.980	-5.120
STATE*	-0.4877	0.085	-5.763	0.000	-0.653	-0.322

Kappa 1 - Logistic Regression

Dep. Variable:	ENROLLMENT_STATUS	**No. Observations:**	47161
Model:	Logit	**Df Residuals:**	47151
Method:	MLE	**Df Model:**	9
Date:	Wed, 02 May 2018	**Pseudo R-squ.:**	0.1429
Time:	20:24:07	**Log-Likelihood:**	-22540.
converged:	True	**LL-Null:**	-26299.

	Coef	**Std Err**	**z**	**P>|z|**	**[0.025**	**0.975]**
DIST	0.0001	3.79e-05	2.893	0.004	3.54e-05	0.000
HIGH_SCHOOL_GPA	-0.0920	0.028	-3.247	0.001	-0.148	-0.036
INTERESTED_IN_CAMPUS_HOUSING	0.1810	0.085	2.131	0.033	0.015	0.348
MAX_TEST_SCORE*	-0.0019	9.45e-05	-19.665	0.000	-0.002	-0.002
SEX	0.1511	0.024	6.260	0.000	0.104	0.198
INST_AID*	3.346e-05	4.31e-06	7.769	0.000	2.5e-05	4.19e-05
VISIT*	5.3205	0.082	64.892	0.000	5.160	5.481
CITY*	15.2709	1.372	11.133	0.000	12.582	17.959
STATE*	2.1365	0.074	28.980	0.000	1.992	2.281
MAJOR*	0.2803	0.140	2.008	0.045	0.007	0.554