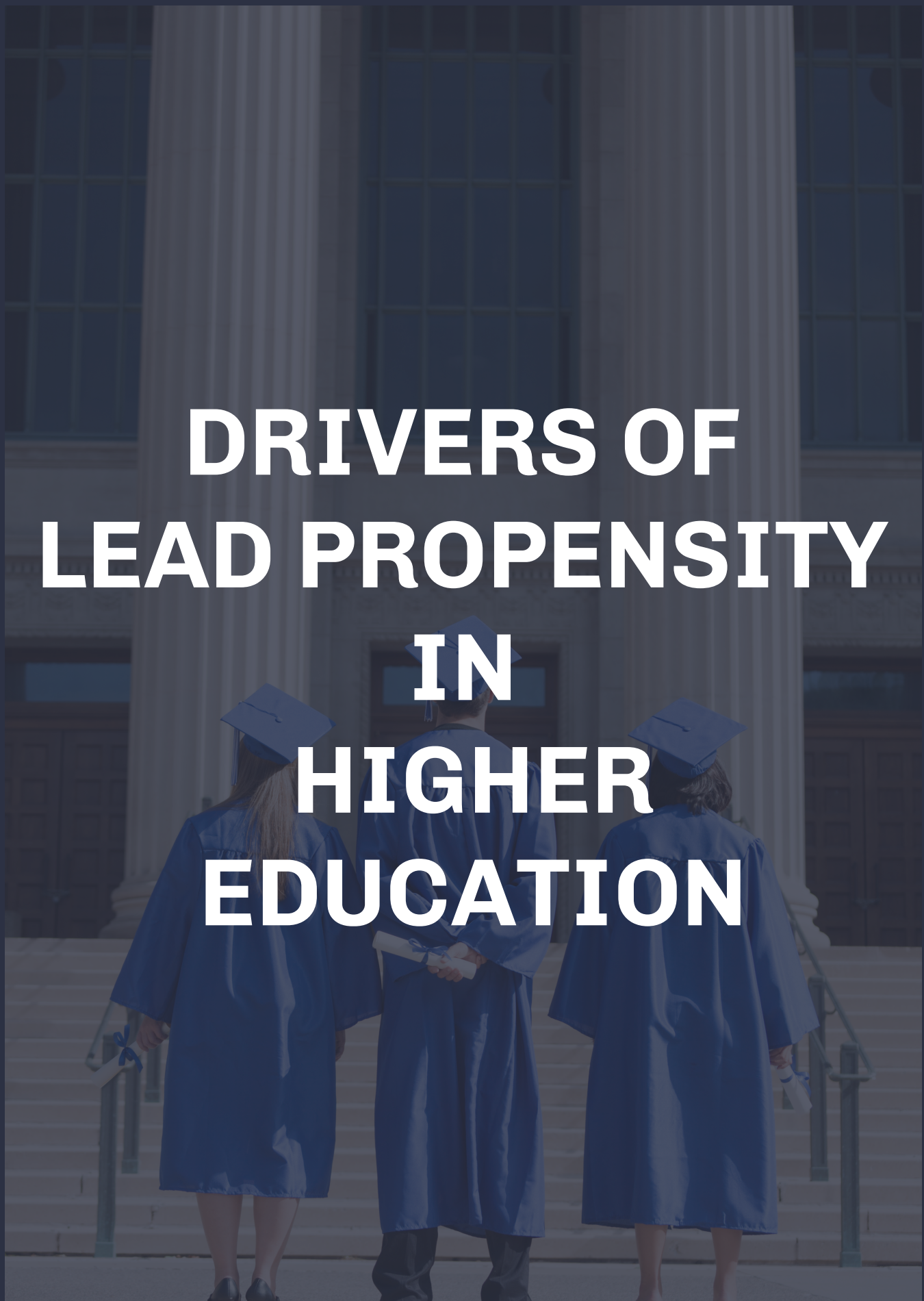

DRIVERS OF LEAD PROPENSITY IN HIGHER EDUCATION

atrium

Drivers of Lead Propensity in Higher Education

Corey Maxedon

Last updated on April 15, 2021

Contents

Introduction

The market for business students is rapidly growing every year. The U.S. Bureau of Labor Statistics cited nearly half a million new jobs in 2015 in management alone (Business Careers). This trend is echoed throughout the world, and students want to prepare themselves well for this business management future. This is why our client in the Shailesh J. Mehta School of Management, Mumbai (SJMSOM) wants to be prepared for this growth. The school of management has a need for a robust lead scoring model in order to make the most out of their available resources when targeting students to take a specific Master of Business Administration level course.

I have been tasked with the creation of this model. I have created two potential models to create a solution for SJMSOM. The first is a logistic regression model using various predictors presented in the student application to take the course. A second model uses a random forest approach for lead scoring. I split the data into a training and testing set for cross validation and model accuracy comparison.

After displaying the results and discussing various features, I concluded that the logistic model would provide the best solution for SJMSOM. It gives a good balance between prediction accuracy and inference of effects. I found the strongest predictors to be a thoroughly completed application with the opt-in for a free interview book. If a student completely filled out the application and opted-in to the interview, they will have a predicted propensity to take the course of 99.66%, leaving all other variables at their baseline value. This is clearly a key aspect for the university to keep track of. Further analysis was completed in Salesforce as well as a final model deployment. This will be used to provide these insights to the sales team in an intuitive, efficient way.

Data

This dataset contains records about students, or leads, wishing to take a class at SJMSOM. There are 9,240 students in the collected data each with up to 49 variables pertaining to a range of details such as demographic and interest information. Most students (6,530) have taken the Management Aptitude Test (MAT), which is a common entrance exam in India for management students. This course is a required course for Masters in Business Administration (MBA) students as many students (5,204) come from a management or business-focused backgrounds. On the other hand, students of all majors can take this course. The goal of this dataset is to determine a lead propensity score using the variable, Is Won?. A lead for the school of management can be defined as a student who has applied for the course. The school's sales team then targets this student and helps them through the sales process with lead stages. The student ultimately either takes the course or decides not to at some point in the sales process; therefore, the student is either "won" or "lost", hence, Is Won?

This data will be used to create a model for new lead propensity scoring. This will help provide insight for the school of management into where their resources will be best spent. There are several marketing efforts and contact campaigns in place, but there is currently no way to know how successful each method is.

Before data inspection can begin, I must take out any observations that are still "in-flight", or in progress. These students are still in the process of signing up for the class and have a status of being "lost", or a status of not being signed up for the class, until a final decision is made (Appendix 2.1). Once these observations are removed, there are still 7,554 students remaining in the dataset. Now, missing value analysis can begin.

Missing Value Analysis

Missingness of data is a major problem in this data. Appendix 1.1 shows how only 21 variables have less than 20%, or about 1,511 observations, of missing values. Even worse, 18 variables have more than 60% of data missing. I engineered a profile strength variable to retain some information from 16 of the variables, while I dropped other columns completely.

First, the new profile strength variable is a way of scoring a student's application on completeness. If a student did not care to take the time to properly complete the application for the class, my hypothesis is the lead will convert less frequently than a more complete application. This provides a way to gain some insight from the variables with extreme amounts of missing data. Table 1 includes all variables I selected to be included in this new variable.

Table 1. Profile Strength Components

Variables in Profile Strength	
Company	Entrance Test
Country	Expectations
Concerns	How Did You Hear About Us?
Awareness Rating	Joining Timeframe
Course Interested	Last Degree
Age	Lead Profile
Current Occupation	Mobile Number
Current Profession	Reason Seeking Adv. Degree

For a single record, a complete application would receive a 16 (one point for each complete variable) while an application with all missing variables would receive a score of zero. This means I did not weight any specific variable any differently than the rest. Figure 1 shows how much the profile strength distribution changes based on a lead being won or lost.

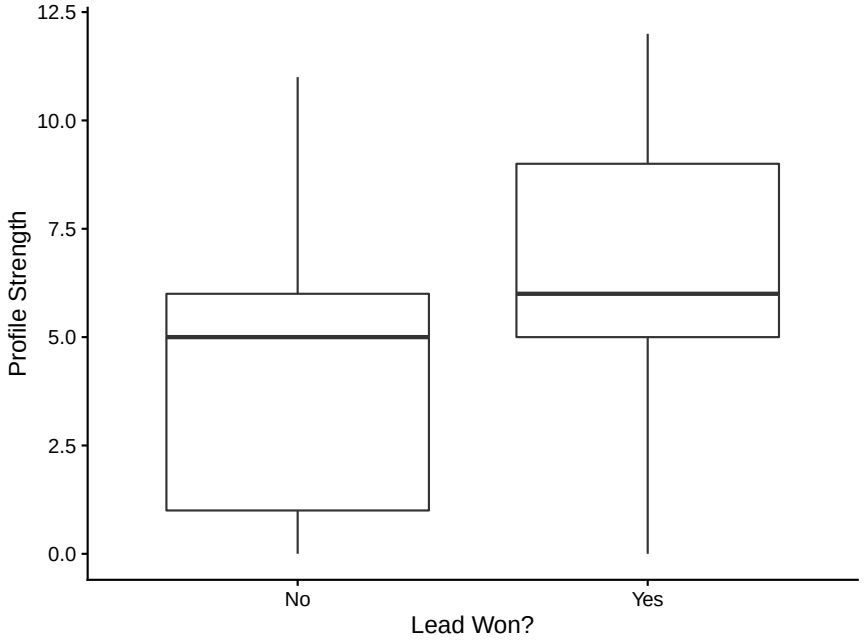

Figure 1: Boxplot of Profile Strength.

Now, I am still left with columns of little value due to missingness. I took a few variables and labeled the missing values as "Not Listed." This may have a potentially interesting effect for specialization, industry, and city with 30.2%, 21.9%, and 34.2% of missingness, respectively. I took this action since it is reasonable to believe some students may not have a specialization, industry, and city to associate themselves with. There was no "Not applicable" option on the application.

For several other variables, the missingness was very high, and I was not able to spin it in a way to gain insight. These variables are included in Table 2. A few other variables also had to be removed for underlying

reasons beyond missingness.

- Is Lost? is the opposite of Is Won? so it was removed
- Lead ID Number is an identification variable which has no explanatory benefit
- Opt-in Course Updates was "No" for all students so no insight could be taken from that except that course updates are not a useful feature in the future
- Last Activity was the same as Last Activity (1), but the university had re-factored it to summarize several levels into a factor level called "Modified." Therefore, it made sense to only keep Last Activity (1) which will be referred to as Last Activity from now on.

Table 2. Variables with High Missingness

Variables with Over 40% Missing Values	
Landing Page ID	Activity Score
Source Campaign	Activity Index
Source Medium	Profile Index
Source Content	Lead Quality
Website	Profile Score
Number of Followup Calls	Lead Profile
Mgmt. Course Match Index	Lead Grade

Tags had a large portion of missingness (~30%), and there were several unrelated factor levels. Tags is more of a notes section specific to each student. I had to remove Engagement Score and Lead Score as these were past, futile attempts at scoring candidates. Last, Lead Stage was removed due to mutually exclusive stages (Appendix 2.2). This means depending on whether a student was won or lost, there was a unique set of stages a student could be listed under. Lead stage is more of a view on where the student is in the sales process. At terminal stages, these values would represent Is Won? status uniquely. The inclusion of this variable would result in a model that is overfit. Further research could involve a time series forecasting approach to lead propensity at each stage. This data is not suited for that analysis though.

After the removal of all of these variables, the dataset is left with the 13 variables in Table 3. Is Won? is the variable I wish to predict. Now, I can look into regrouping some of the categorical variables with more than about 5 factor levels.

Table 3. Variable Description

Variables to be Analyzed

Variable	Type	Explanation
Is Won?	Categorical	Lead is won. "Yes" or "No".
City	Categorical	Place of origin.
Profile Strength	Numeric	Measure of application completeness.
Opt-In Free Interview Book	Categorical	Lead wants interview book. "Yes" or "No".
Lead Origin	Categorical	Describes how the university acquired the lead.
Lead Source	Categorical	Method student took to find and complete application.
Last Activity	Categorical	Last activity completed with application.
Industry	Categorical	Industry associated to profession.
Specialization	Categorical	Specialization of profession.
Total Views	Numeric	Total views of the application.
Total Visits	Numeric	Total visits to the application.
Views per Visit	Numeric	Number of views per visit to the application.
Avg. of Time per Visit	Numeric	Average time spent on application per visit.

Variable Transformation

In order to model the data correctly without overfitting, or producing an inaccurate, over-confident prediction, I find it necessary to reduce the number of factor levels for several variables. Appendix 1.2 shows all variables I wish to re-factor and shows the make-up of the proposed groupings. Appendix 1.3 displays the distribution of the original levels versus the new levels.

For each variable, I tried to make less than five groups that made intuitive sense. For example, Specialization originally had 19 different factor levels. The main insight I want to take away from Specialization is the difference between management, not management, and unlisted. As I said, it is important to keep the missing values through a “Not Listed” level. The student may not have a career yet so it is unreasonable to expect all missing values were done so at random. Now, all data cleaning has been done, but the data still has 167 observations containing missing values. Out of the 7,554 observations left, this few of records will not produce a major impact on the predictive ability of the model. Besides, I have no reason to believe these values are missing at random so it is best to practice listwise deletion in this case. This leaves me with 7,387 observations with 13 important variables each.

Distribution of Variables

With all data now in order, the distribution and summary of the variables can be observed. Table 5 includes the categorical variables while Table 6 displays the distribution of quantitative variables. The categorical variables from the table alone do not allow me to gain much insight. A visual representation will allow a fuller view of the underlying relationships.

Table 5. Categorical Variables

Variable	Factor Level	Count	Percent
Is Won?	No	3,955	53.5
	Yes	3,432	46.5
Specialization	Management	4,122	55.8
	Not listed	2,183	29.6
	Not Management	1,082	14.6
City	Mumbai	2,805	38.0
	Not listed	2,496	33.8
	Other Cities	2,086	28.2
Industry	Consumer Durables	5,668	76.7
	Not Consumer Durables	73	1.0
	Not listed	1,646	22.3
Last Activity	Email/Chat	2,613	35.4
	Phone	2,166	29.3
	Unreachable	105	1.4
	Website	2,503	33.9
Lead Origin	API	2,501	33.9
	Landing Page	4,279	57.9
	Manual Import	607	8.2
Lead Source	Direct Traffic	2,044	27.7
	Google	2,638	35.7
	Other	1,662	22.5
	Other Search Engine	1,043	14.1
Free Interview Book?	No	4,943	66.9
	Yes	2,444	33.1

n = 7,387

There is one interesting detail about the quantitative variables Total Views, Total Visits, Views per Visit, and Avg. Time per Visit. The distribution of these variables is wildly variable and large. The max value of each is extremely far from the median. Another thing to note is each variable includes zero. As I begin the data analysis and think about possible transformation, this will be important.

Table 6. Numeric Variables

	Profile Strength	Total Views	Total Visits	Views per Visit	Avg. Time per Visit
Min.	0.0	0.0	0.0	0.0	0.0
1st Qu.	5.0	4.0	2.0	1.3	3.0
Median	5.0	9.0	3.0	2.0	77.0
Mean	5.3	14.6	3.8	2.6	132.7
3rd Qu.	7.0	16.0	5.0	4.0	194.7
Max.	12.0	946.1	251.0	24.0	1199.0

Initial Data Analysis (IDA)

The data analysis can now begin with the well defined variables I uncovered before. The first thing to visualize would be the possible transformation of some of the quantitative variables mentioned. It is important to perform the transform now as I may want to see relationships with the transformed variables later. Figure 2 displays the distribution of each of the variables. As I suspected, the distributions are significantly

asymmetric with severe outliers. It appears combining Total Views and Total Visits into one variable, Views per Visit, reduces the spread of the data by a large margin.

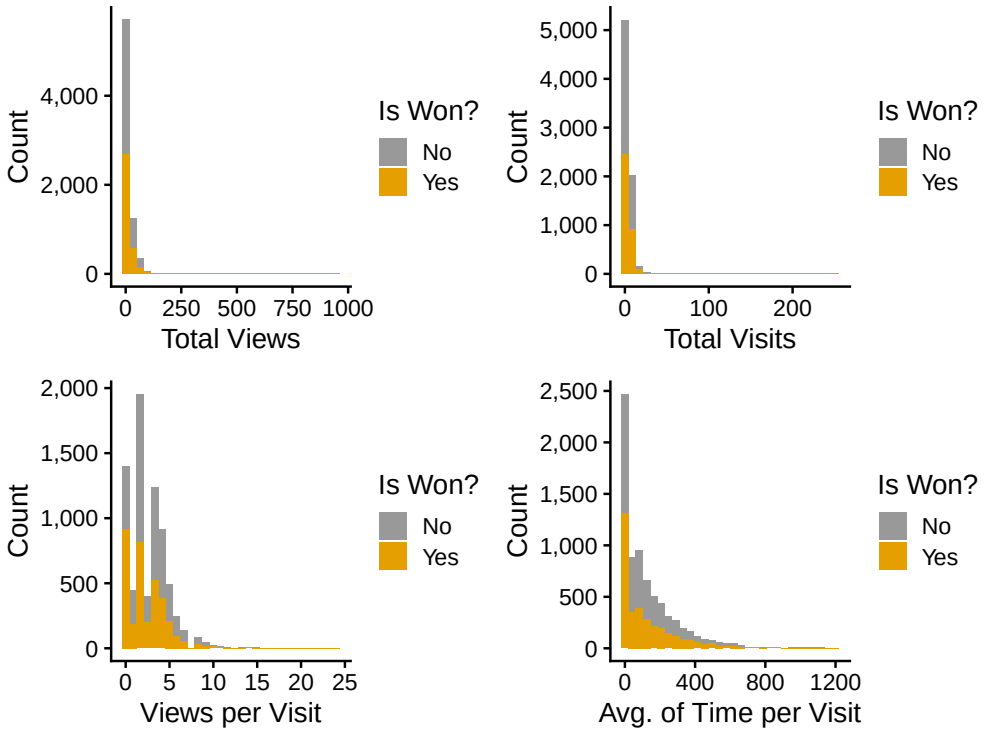

Figure 2: Distribution of Visits and Views.

It is reasonable to continue analysis with only Views per Visit and Avg. of Time per Visit, but transformation must be analyzed as the distributions are still not symmetric. Figure 3 shows the difference in each original variable versus a natural logarithmic (log) transformation. As noted, these variables do include zero so a negligible value of 0.1 was added to each observation in order to perform the log transformation. The log transformation shows a much more symmetric distribution than before. There is a large amount of zero values so the only skew now is caused by these values. This comparison shows support to continue with the log transformation on these variables. This figure also shows a relationship between each variable and Is Won?.

Next, I looked at the relationship between Is Won?, Industry, Free Interview Book?, and Specialization (Figure 4). The diagonal shows the distribution of each individual variable while other mosaic plots show the relationships between variables. The top row gives a feel for the relationships between each variable and the outcome variable, Is Won?. A change in box size going from Is Won?:“Yes” to Is Won?:“No” shows a clear relationship. It is evident that each variable is related in some way to the outcome variable.

There is a clear relationship between Last Activity and Is Won? (Figure 5). A student’s last activity has varying effects on lead propensity to close. This will be an important effect to examine in the modeling process.

Next, I look at the potential for interactions between variables. I hypothesized that Lead Origin and Lead Source would have an interaction when trying to explain the outcome. The reasoning would be that some leads may convert more often when coming from a specific origin that associates to a specific source whereas another source from that same origin may not convert as well. This would present a need for an interaction in the model. Figure 6 shows this thought visually without any evidence of interaction. If there were an interaction, the distribution of Lead Source within each origin would change drastically when the lead goes from “Won” to “Lost”. What the figure shows is no change in distribution at all which leads me to conclude no significant evidence of an interaction being effective.

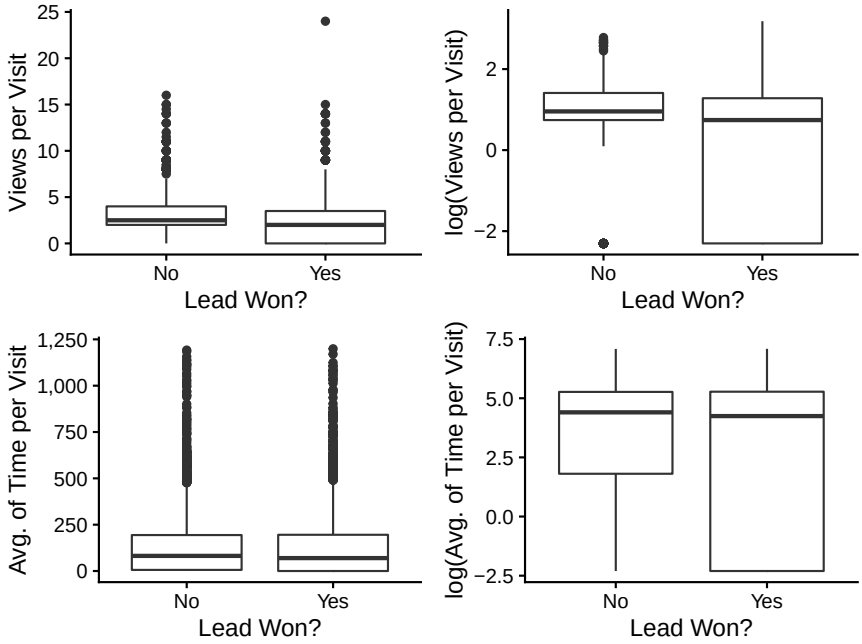

Figure 3: Natural Logarithm (log) Transformation of Views and Visits.

Another potential interaction to look at is that of log(Avg. Time per Visit) and log(Views per Visit). It seems like the relationship between these variables and Is Won? would change depending on the outcome. Figure 7 shows no evidence to support this interaction. The smoother lines (shown for visualization purposes) are unchanged when the outcome is "Won" or "Lost".

The last potential interaction is a between Lead Origin, Lead Source, and log(Views per Visit). It seems that if a person completes the application on a 3rd party site (API) then the views per visit would have a lower effect on explaining the outcome, but if the application is completed through the home website, the views per visit could explain a much larger portion of the effect. Figure 8 represents this interaction with Is Won?. What I see is very little change in the distribution in source by origin among different levels of log(Views per Visit) when the outcome changes between "Won" and "Lost". This information does not support the interaction.

Overall, I found no significant interactions to model, but each individual variable seems to have some relationship which will be more clearly identified in the modeling phase.

Methods

With all data processed and important relationships visualized, the lead scoring model can now start to be identified for SJMSOM. The outcome variable in this case is Is Won? which is a binary, categorical variable. This data is most appropriate to be fit with one of two models.

The first model is a logistic regression approach to produce lead propensity based on the features I have identified. This model will be rigorously tested to ensure the model gives accurate predictions on future student leads. A final model will be able to produce lead propensity in terms of a probability of winning the lead. Once a propensity has been found, a threshold for predicted "Win"/"Loss" can be made in order to give a definitive prediction. On top of the predictive insight, the model will allow SJMSOM to know how each variable and factor impacts propensity.

A second model will invoke a decision tree based method called a random forest. This technique is at the forefront of predictive modeling, but one downside is the loss of the same level of interpretability that the

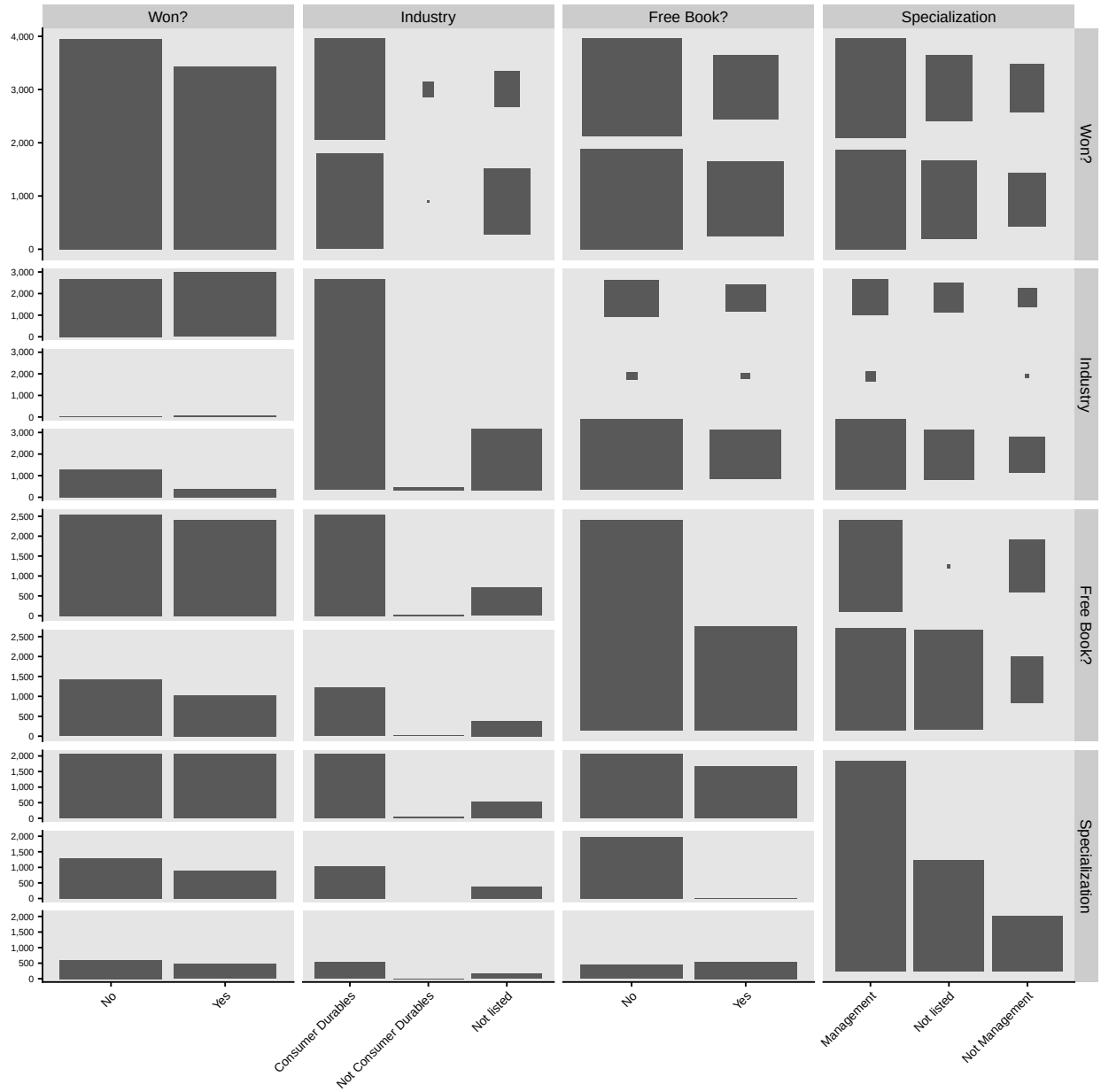

Figure 4: Distribution of Categorical Variables. The diagonal shows the distribution of each individual variable while other mosaic plots show the relationships between variables. The top row gives a feel for the relationships between each variable and the outcome variable, Is Won?. A change in box size going from Is Won?:'Yes' to Is Won?:'No' shows a clear relationship.

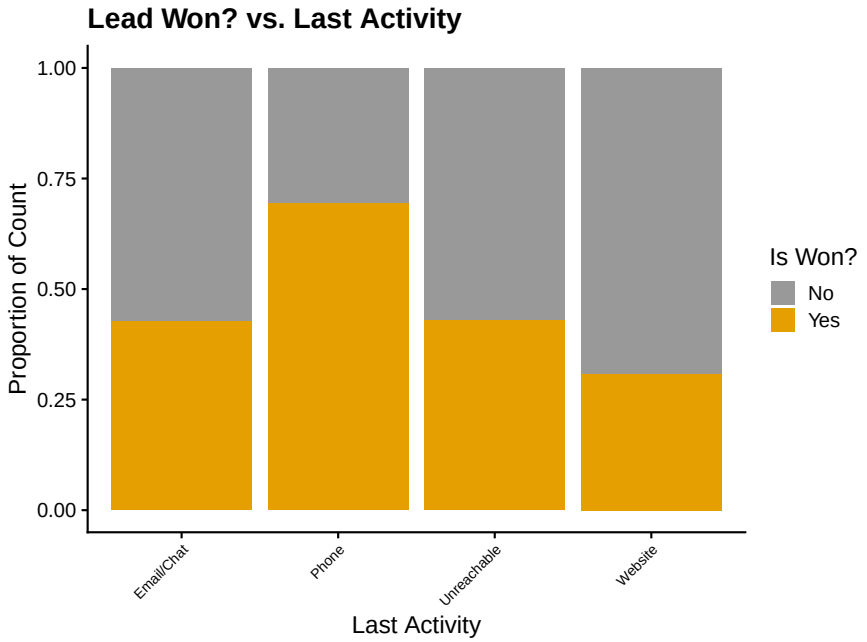

Figure 5: Distribution of Last Activity.

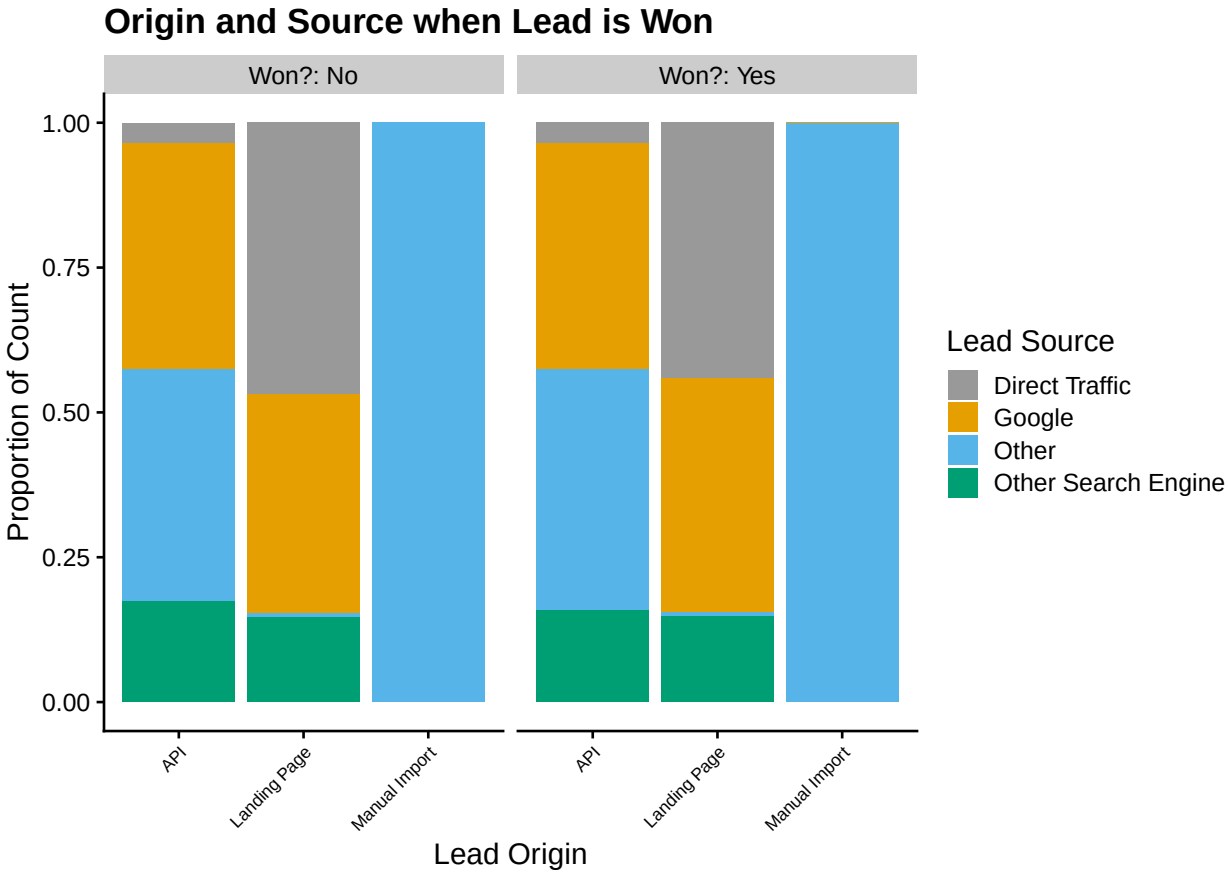

Figure 6: Distribution of lead source within lead origin is unchanged when going from 'Won' to 'Lost'.

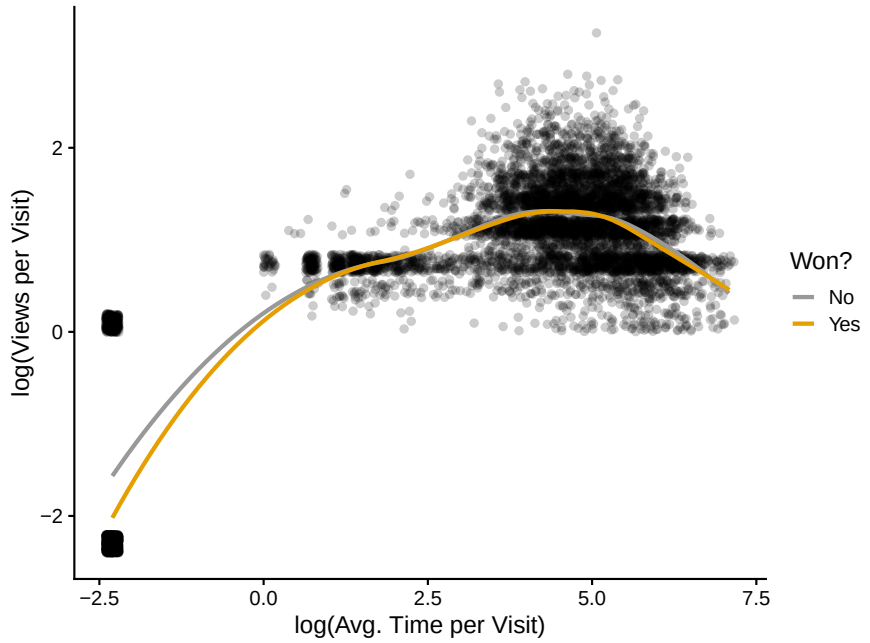

Figure 7: The relationship is identical when going from 'Won' to 'Lost'.

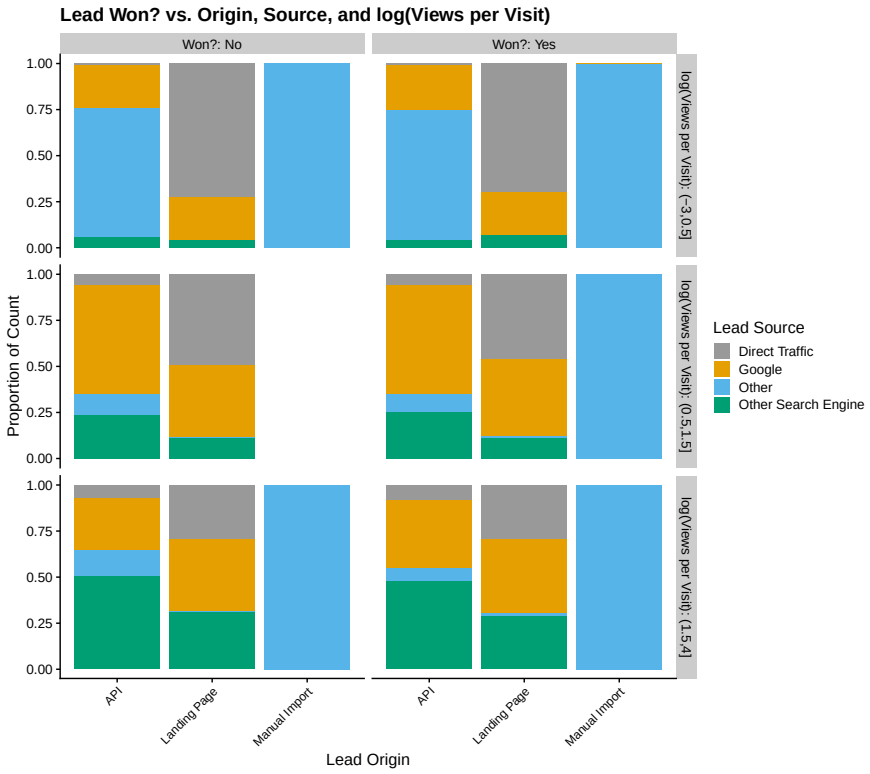

Figure 8: The three-way interaction does not provide a significant difference in distribution when going from 'Won' to 'Lost'.

logistic model provides. This model will produce a propensity score as well as a prediction of "Win"/"Loss", but the impact of each variable and their factors is not as in-depth as the logistic model is able to provide.

With this information in mind, it is important to build both of these models and test their predictive ability against each other. This will be done by splitting the data into a training set with 70% of the original data and a testing set with 30% of the data. This will allow SJMSOM to be confident in the output from these models.

Model 1: Logistic Regression for Lead Scoring

The first model I created was a logistic model of lead propensity which used all important variables discovered in IDA. By performing model selection using the Akaike information criterion (AIC), a model-fitting statistic which penalizes both predictive error and excessive numbers of model parameters, I conclude the model with the lowest AIC value is one that includes all variables identified except for Lead Source. The AIC was minimized by removing that variable alone. The final residual deviance was 5,347 which a major improvement over the null deviance of 7,144.

Next, I performed further diagnostic checks to test if the underlying assumptions of the model were valid. The distribution of residuals in the residual plot seen in Figure 9 demonstrate no patterns aside from what is to be expected for a logistic regression. I use this plot to ensure the residuals do not reveal outliers or trends aside from the grouping patterns seen. Likewise, the half-normal plot in Figure 10) also shows no assumptions are broken; I see one point with a significant jump in the trend of the graphed data, but point 1,581 had above average log(Avg. Time per Visit) (5.17 vs. 2.98 average). It is unreasonable to take out this outlier. Some candidates may have higher than expected time viewing the application. The main assumptions for the logistic model have been met, and I can continue with an estimate of the model's predictive ability.

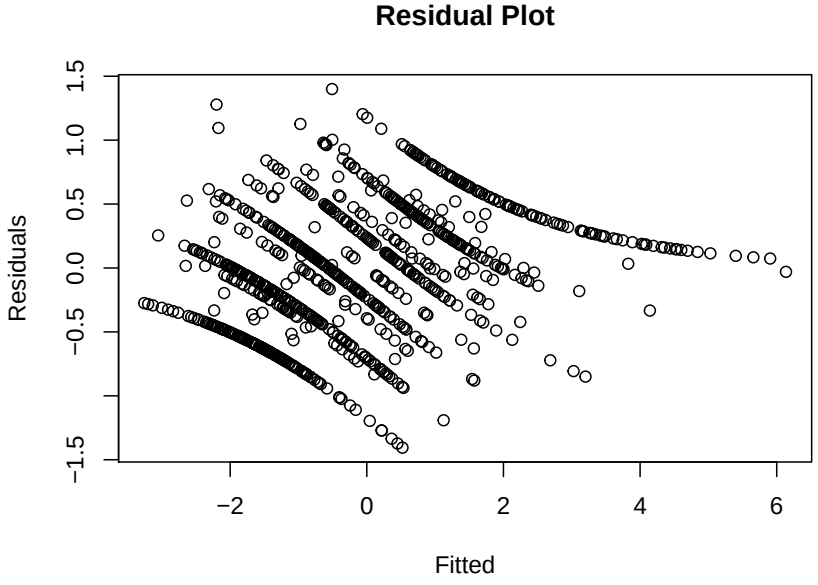

Figure 9: Logistic Model Residual Plot

I looked at the receiver operating characteristic (ROC) curve to evaluate the predictive ability of the logistic model (Figure 11). The model provides an Area Under the Curve (AUC) of 0.820. A model with perfect predictive ability would have an AUC of 1, while a model with the worst possible predictive ability would have an AUC of 0.5, equivalent to a random chance of guessing correctly. An AUC of 0.820 is moderately predictive and verifies this model is appropriate to use for further prediction.

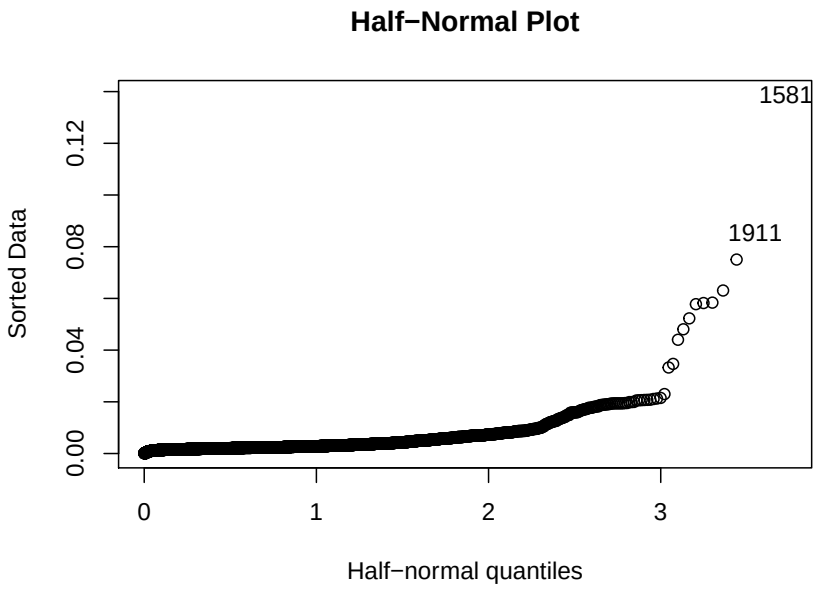

Figure 10: Logistic Model Half-Normal Plot

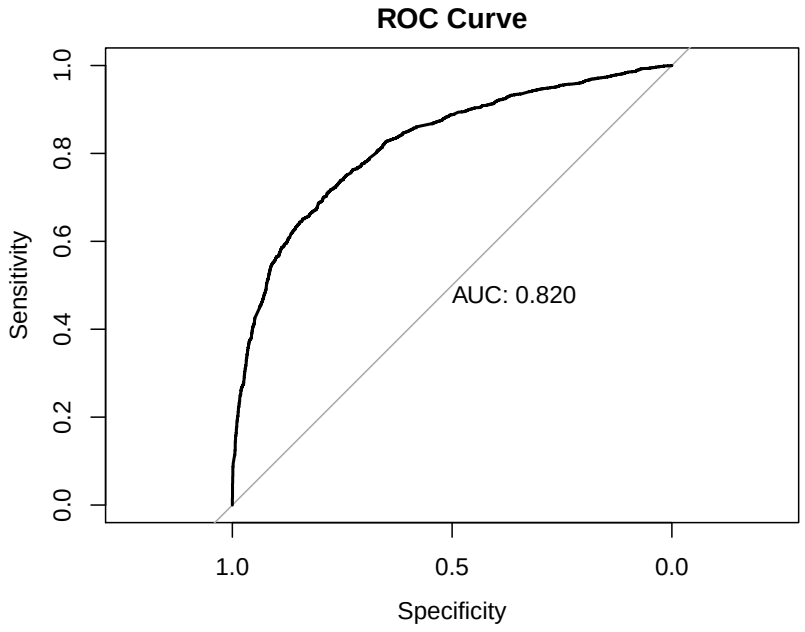

Figure 11: ROC curve for the Logistic Model.

A final step before being able to compare predictive accuracy is finding the optimal threshold for the break between "Yes" and "No" for Is Won?. The logistic model produces propensity scores so the cutoff point needs to be found. In order to find this point, I made a grid of possible thresholds and used it to predict against the data in the training set. Figure 12 shows the optimal threshold for peak predictive ability on the training data. This figure completes a grid search for threshold on the x-axis and plots the corresponding accuracy on the y-axis. I define optimal threshold as the highest point in this curve. The threshold turns out to be 55% which will be used in model comparison testing. This means a lead with a propensity over 55% will be categorized as a "Win".

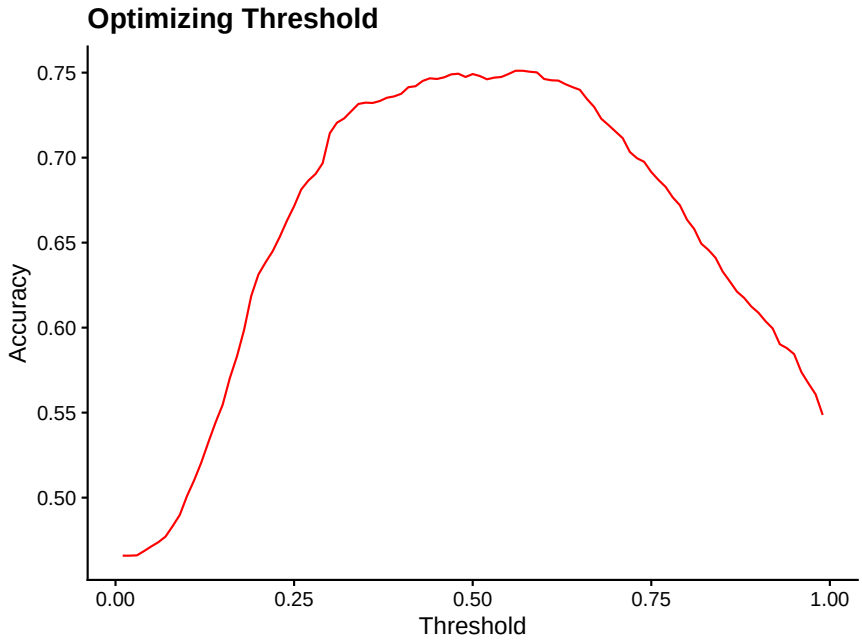

Figure 12: Finding maximum accuracy for logistic model.

Model 2: Decision Tree Building for Lead Scoring

Lead scoring is a perfect candidate for a categorical tree. I lose the quantitative inference capability about specific features that the last model gives, but the prediction ability is better in many cases. Decision trees are simple. Figure 13 gives an example of a decision tree. Each bubble gives three values: a word, a number, and a percentage. The word is the predicted outcome at that specific point, or node, in the tree. The number is the propensity score at that node. The percent is the amount of data in the training set that made it to that specific node. Below each bubble is a decision. If the logical expression is true, the record continues to the left in the tree. Once a record reaches a final point with no additional decisions, a final prediction is given.

A random forest scales this idea up. Instead of producing one tree, hundreds of trees are developed in this same manner. Two differences also apply when building the forest. Each tree is built using a random subset of the training data with replacement. Also, each decision node can only pick a select number of variables to use for the decision. In the single tree, the decision was based on any variable that provided the greatest predictive power. Now, only a select number, three for example, of possible variables are allowed to be pulled from a hat for use in building the decision.

A key part of building a random forest is hyper-parameter tuning. This is a fancy term for saying I want to pick an optimal number when I choose how many variables I can pull from the hat at each node. This tuning can be done by using a grid search. I build models using different values for the number of variables to be pulled at each node. Figure 14 displays the accuracy of each model in the grid search. The search could

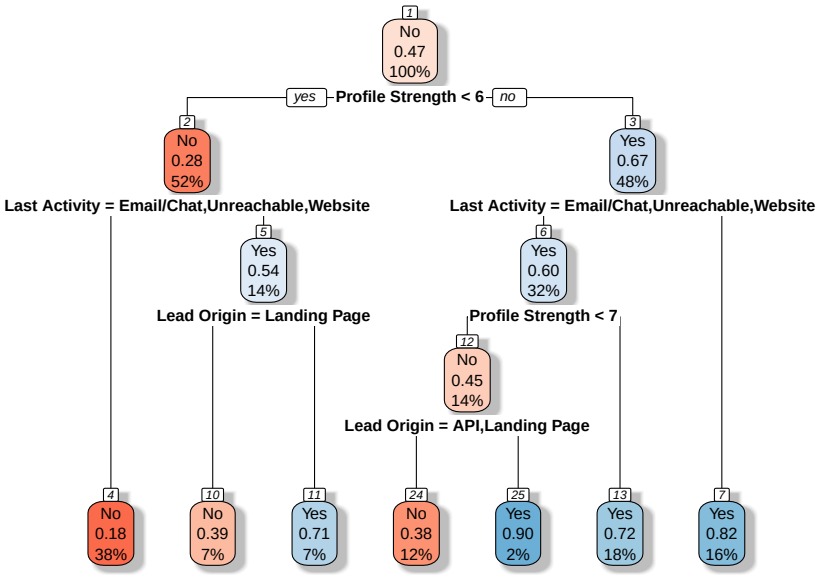

Figure 13: Example of a decision tree using this data.

have been numbers between 1 through 11 since I have 11 total variables included in the model. The plot shows a clear maximum point at 5. This means each node will be able to choose from 5 randomly selected variables when trying to develop the optimal predictive decision.

Comparison of Model Performance

Now, each model has been developed using training data. The university gave me ample data to be able to provide a look at predictive power on a test data set. The logistic model was able to predict on the training data with accuracy of 74.92% while the random forest was able to predict the training data with 87.37% accuracy. The real test will be the comparison between each on the test data.

Table 7 and 8 are error tables for the logistic and random forest models, respectively. Overall, the logistic model was able to predict accurately 75.05% of the time while the random forest model was able to predict correctly 77.35% of the time. This is very close. If predictive accuracy was the only concern, the random forest would be desired, but the university also needs to know specific contributors in order to drive higher lead conversion.

Table 7. Logistic Model Test Set Error

Is Won?	% Error
No	18.1056
Yes	32.9423

Table 8. Random Forest Model Test Set Error

Is Won?	% Error
No	18.3571
Yes	27.6637

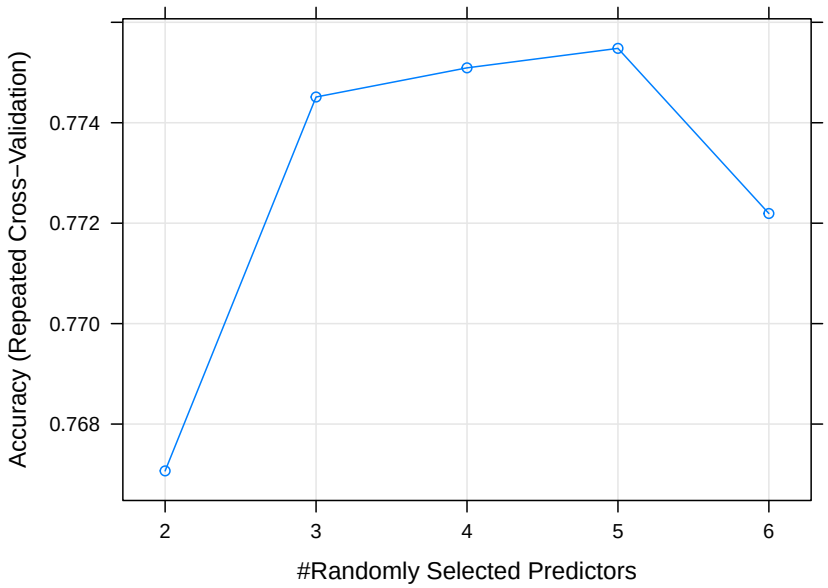

Figure 14: Finding the optimal number of possible variables to be tried at each node.

Discussion

In the logistic model, I started with all variables as predictors with lead winning as the response. After model selection, I removed Lead Source. This decision was based on minimizing stepwise AIC. The model output is seen in Table 9. Each estimate has an associated standard error in order to see how variable a coefficient can be when compared to magnitude.

Table 9. Logistic Model Coefficients

Variables		Estimate	Std. Error
Intercept		-3.89	0.25
Specialization	Management	-	-
	Not listed	-0.73	0.16
	Not Management	-0.24	0.10
City	Mumbai	-	-
	Not listed	0.67	0.17
	Other Cities	0.12	0.08
Industry	Consumer Durables	-	-
	Not Consumer Durables	1.36	0.77
	Not listed	2.31	0.17
Profile Strength		0.57	0.03
Last Activity	Email/Chat	-	-
	Phone	1.44	0.08
	Unreachable	0.25	0.27
	Website	-0.38	0.08
Lead Origin	API	-	-
	Landing Page	-1.05	0.13
	Manual Import	1.61	0.20
Free Interview Book?	No	-	-
	Yes	1.02	0.10
log(Views per Visit)		-0.33	0.06
log(Avg. Time per Visit)		0.08	0.02

Note: Coefficients given in terms of log(Odds Ratio) = log(OR) = log(Probability of Success / Probability of Failure). Positive values mean higher propensity to "Win" while negative values mean lower propensity to "Win". log(OR) of 0 translates to a propensity of 50%.

Based on the log of the odds ratio (OR) of the logistic model I obtained (Table 9), I have evidence to conclude that, compared to a student with a specialization in management, from Mumbai, in the consumer durables industry, with a last activity of email, with a lead origin from a third party, and who did not opt in to the free interview book (the baseline with $log(OR) = -3.89$ or propensity of 2.0%), the potential student with optimal propensity has the following characteristics, holding other factors constant:

- Specialization in Management (baseline)
- City is Not Listed (change of $log(OR) = 0.67$)
- Industry is Not List (change of $log(OR) = 2.31$)
- Profile Strength is high (change of $log(OR) = 0.57$ per unit)
- Last Activity is Phone (change of $log(OR) = 1.44$)
- Lead Origin is API (baseline)
- Opt-in to the free interview book (change of $log(OR) = 1.02$)
- Have low log(Views per Visit) (change of $log(OR) = -0.33$ per unit)
- Have high log(Avg. Time per Visit) (change of $log(OR) = 0.08$ per unit)

A student with these characteristics would have a propensity to take the class of 99.9%. Realistically, based on this model, I can say industry, profile strength, activity, origin, and free book decision are key indicators of success. If the industry is not listed, this has a high positive impact on lead propensity. Also, as hypothesized, a student putting focus in creating a complete application and taking the initiative to opt-in to the free interview book has possibly the highest indication of taking the class. The profile strength

variable has a maximum score of 16. If a student completely filled out the application (16) and opted into the book, this would created a log(OR) increase of 9.57 from the baseline candidate. On this alone, the propensity would shift from 2% at the baseline to 99.66% propensity to take the course. This is clearly a key aspect for the university to keep track of.

Now, I would like to compare these results to that of the random forest. Instead of coefficients, this model gives importance values to the most important effects. Table 10 shows the sorted list of the most important variables and factor levels. Unlike the logistic model, I am unable to see the exact magnitude change when I look at a specific variable. On the other hand, based on relative importance score, I can see some parallels, such as how important profile strength is. Profile strength is by far the highest indicator of lead propensity. Surprisingly, log(Avg. Time per Visit) and log(Views per Visit) are much more important in this model than the logistic model. Unfortunately, I cannot see the direction of how each variable is related to Is Won?, negative or positive.

Table 10. Random Forest Importance

Variable	Importance
Profile Strength	464.15
log(Avg. Time per Visit)	314.21
log(Views per Visit)	195.46
Last Activity: Phone	171.87
Lead Origin: Manual Import	81.96
Last Activity: Website	71.09
Industry: Not Listed	59.24
Lead Origin: Landing Page	41.83
Free Book?: Yes	41.65
City: Other Cities	38.59
Lead Source: Google	38.53
Specialization: Not Listed	31.82
Specialization: Not Management	30.67
Lead Source: Other Search Engin	26.91
Lead Source: Other	26.43
City: Not Listed	25.41
Last Activity: Unreachable	8.76
Industry: Not Consumer Durables	5.08

Overall, the logistic model is the most appropriate for the Shailesh J. Mehta School of Management, Mumbai. This model will not only give high quality predictions, but also allow insights into exactly how individual factor levels related to lead propensity.

Salesforce Implementation

With a model decided on, it is time to deploy this model. A model by itself is useful and provides insight, but a model that provides real time insight for each individual salesperson can turn out to be invaluable. This is exactly the capability Salesforce Einstein Discovery (SED) provides.

SED tries to be an automatic machine learning (autoML) platform. After all, the target audience of this platform is the typical businessperson, not necessarily a data scientist or statistician. As such, some desired capabilities typically performed in a traditional analysis, such as model selection, do not have the ability for user input. The platform does provide automatic 5-fold Cross Validation as well as a moderate variety of model capability to provide solutions for the most common business problems. These models include:

- Generalized Linear Model (GLM)
- Gradient Boost Machine (GBM)
- Extreme Gradient Boost (XGBoost)
- Random Forest
- Model Tournament

As well as not providing the capability for user input, the service also makes some interesting default decisions, such as lasso regression for GLM, two-way interactions between all predictors, and mandatory binning of all discrete and continuous variables. This is not all bad. Business problems have several moving parts. A model could contain tens of possible predictors. The lasso regression will provide regularization and bring some coefficients to zero in order to help the most predictive variables stand out while also preventing overfitting to the training data.

SED also brings evaluation metrics, exploratory data analysis, and an interactive prediction analysis together in an easy-to-use way. Some of these metrics would make a statistician cringe, such as reporting R-squared on a logistic model, but one thing both types of end users can agree on is the ease of model deployment.

Since an independent analysis gave me the most flexibility in determining an appropriate model, I completed the data engineering, visualization, and modeling all before ever even opening Einstein Discovery. SED is not an autoML for the level of analysis I complete as a data scientist. It is merely a tool to deploy robust models on large enterprise scale. This will provide real time insights and analyses every salesperson can appreciate.

I was able to provide this detail to SJMSOM salespeople through sorted list views and side card views of potential drivers. The list view will provide a list of all student leads and sort by lead score. This will allow a salesperson to either avoid low scoring leads and focus on converting higher scoring leads, or dig into why a lead is scoring so low with the hopes of a discovery of a systematic change that could help the entire sales team. If a salesperson wants to dive into a specific record, they have the ability to drill in and gain feedback and recommendations on what is working as well as what is not for that particular student. I was able to build and deploy all of this through the power of Salesforce Einstein Discovery.

Conclusion

The trend toward management-related positions is growing throughout the world, and students want to prepare themselves well for this business management future (Business Careers). Our client, the Shailesh J. Mehta School of Management, Mumbai, wants to be prepared for this growth. The school of management has a need for a robust lead scoring model in order to make the most out of their available resources when targeting students to take a specific Master of Business Administration level course.

I have created two models to create a solution for SJMSOM. The first is a logistic regression model using various predictors presented in the student application to take the course. A second model uses a random forest approach for lead scoring. I split the data into a training and testing set for cross validation and model accuracy comparison. After displaying the results and discussing various features, I conclude that the logistic model would provide the best solution for SJMSOM.

It gives a good balance between prediction accuracy and inference of effects. Based on the log of the odds ratio (OR) of the logistic model I obtained (Table 9), I have evidence to conclude that, compared to a student with a specialization in management, from Mumbai, in the consumer durables industry, with a last activity of email, with a lead origin from a third party, and who did not opt in to the free interview book (the baseline with $log(OR) = -3.89$ or propensity of 2.0%), the potential student with optimal propensity has the following characteristics, holding other factors constant:

- Specialization in Management (baseline)
- City is Not Listed (change of $log(OR) = 0.67$)
- Industry is Not List (change of $log(OR) = 2.31$)

- Profile Strength is high (change of $log(OR) = 0.57$ per unit)
- Last Activity is Phone (change of $log(OR) = 1.44$)
- Lead Origin is API (baseline)
- Opt-in to the free interview book (change of $log(OR) = 1.02$)
- Have low log(Views per Visit) (change of $log(OR) = -0.33$ per unit)
- Have high log(Avg. Time per Visit) (change of $log(OR) = 0.08$ per unit)

I found the strongest predictors to be a thoroughly completed application (Profile Strength) with the opt-in for a free interview book. If a student completely filled out the application and opted-in to the interview, they will have a predicted propensity to take the course of 99.66%, leaving all other variables at their baseline value. This is clearly a key aspect for the university to keep track of. Further analysis was completed in Salesforce as well as a final model deployment. This will be used to provide these insights to the sales team in an intuitive, efficient way.

Appendix 1. Additional Tables

Appendix 1.1. Missing Data After Filtering In-flight Records

Variable	% Missing	Variable	% Missing
Is Lost?	0.0%	Tags	29.7%
Industry	0.0%	Landing Page ID	43.3%
Engagement Score	0.0%	Profile Score	44.3%
Lead Score	0.0%	Activity Score	44.3%
City	0.0%	Activity Index	44.3%
Is In-flight?	0.0%	Profile Index	44.3%
Is Won?	0.0%	Lead Quality	46.9%
Last Activity (1)	0.0%	Source Campaign	64.3%
Lead ID Number	0.0%	Source Medium	64.3%
Lead Origin	0.0%	Source Content	65.3%
Lead Stage	0.0%	Lead Profile	72.7%
Opt-in Course Updates	0.0%	How Did You Hear About Us?	74.5%
Opt-In Free Interview Book	0.0%	Awareness Rating	78.8%
Specialization	0.0%	Joining Timeframe	78.9%
Profile Strength	0.0%	Expectations	79.2%
Lead Source	0.5%	Lead Grade	79.7%
Last Activity	1.4%	Number of Followup Calls	80.2%
Total Visits	1.8%	Mgmt. Course Match Index	88.4%
Views per Visit	1.8%	Concerns	89.3%
Total Views	1.8%	Last Degree	93.5%
Avg. of Time per Visit	1.8%	Company	96.0%
Current Occupation	21.8%	Age	97.8%
Country	21.9%	Course Interested	99.0%
Current Profession	21.9%	Mobile Number	99.6%
Reason Seeking Adv. Degree	21.9%	Website	99.7%
Entrance Test	21.9%		

n=7554; after filtering out Is In-flight? = "Yes"

Appendix 1.2. Updated Factor Levels

Variable: Specialization	
Groups	**Original**
Management	Finance Management
	Human Resource Management
	Marketing Management
	Operations Management
	Business Administration
	Supply Chain Management
	IT Projects Management
	Healthcare Management
	Hospitality Management
	Retail Management
Not Management	Banking
	Media and Advertising
	International Business
	Travel and Tourism
	E-Commerce
	Rural and Agribusiness
	E-Business
	Services Excellence
Not Listed	NA's

Variable: Last Activity	
Groups	**Original**
Website	Modified
	Page Visited on Website
	Approached upfront
Email/Chat	Email Opened
	Email Link Clicked
	Olark Chat Conversation
	Email Received
	Resubscribed to emails
Phone	SMS Sent
	Had a Phone Conversation
Unreachable	Unsubscribed
	Unreachable
	Email Bounced
	Email Marked Spam

Variable: Lead Source	
Groups	**Original**
Google	Google
Direct Traffic	Direct Traffic
Other Search Engine	Organic Search
	Bing
Other	Olark Chat
	Reference
	Welingak Website
	Referral Sites
	Facebook
	Click2Call
	Live Chat
	Social Media
	Blog
	EDM
	Pay per Click Ads
	WeLearn
	YouTube
Missing to be Removed (34	NA's

Variable: Industry	
Groups	**Original**
Consumer Durables	Consumer Durables
Not Listed	NA's
Not Consumer Durables	Banking
	Advertising
	Medical
	IT
	Recruitment
	Travel
	Shipping
	Construction
	Export/Import
	Financial Services
	Food Processing
	Industrial Products

Variable: City	
Groups	**Original**
Mumbai	Mumbai
Not Listed	NA's
Other Cities	Thane & Outskirts
	Other Cities
	Other Cities of Maharashtra
	Other Metro Cities
	Tier II Cities

Variable: Lead Origin	
Groups	**Original**
Landing Page	Landing Page Submission
3rd Party	API
Manual Import	Lead Add Form
	Lead Import
	Quick Add Form

Note: NA means missing value.

Appendix 1.3. Proposed Factor Groups

Variable: Specialization	%
NA's	30.2
Finance Management	11.5
Human Resource Management	10.2
Marketing Management	10.2
Operations Management	6.1
Business Administration	5.1
Banking	4.2
Supply Chain Management	4.1
IT Projects Management	3.8
Media and Advertising	2.5
International Business	2.2
Travel and Tourism	2.2
Healthcare Management	2.0
Hospitality Management	1.3
E-Commerce	1.3
Retail Management	1.2
Rural and Agribusiness	0.9
E-Business	0.6
Services Excellence	0.4
Proposed Grouping	**%**
Management	55.4
Not Listed	30.2
Not Management	14.4

Variable: Lead Source	%
Google	34.9
Direct Traffic	27.1
Organic Search	13.8
Olark Chat	12.3
Reference	7.1
Welingak Website	1.9
Referral Sites	1.6
Facebook	0.7
NA's	0.5
Click2Call	0.05
Live Chat	0.03
Social Media	0.03
Bing	0.01
Blog	0.01
EDM	0.01
Pay per Click Ads	0.01
WeLearn	0.03
YouTube	0.01
Proposed Grouping	**%**
Google	34.9
Direct Traffic	27.1
Other	23.8
Other Search Engine	13.8
Missing to be Removed (34)	0.5

Variable: Lead Origin	%
Landing Page Submission	56.7
API	33.1
Lead Add Form	9.5
Lead Import	0.7
Quick Add Form	0.01
Proposed Grouping	**%**
Landing Page	56.7
3rd Party	33.1
Manual Import	10.2

Variable: City	%
Mumbai	37.8
NA's	34.2
Thane & Outskirts	9.0
Other Cities	8.3
Other Cities of Maharashtra	5.5
Other Metro Cities	4.5
Tier II Cities	0.8
Proposed Grouping	**%**
Mumbai	37.8
Not Listed	34.2
Other Cities	28.0

Variable: Last Activity	%
Email Opened	32.2
Modified	31.9
SMS Sent	28.8
Page Visited on Website	3.1
Email Link Clicked	1.6
Olark Chat Conversation	0.8
Unsubscribed	0.6
Unreachable	0.4
Email Bounced	0.4
Had a Phone Conversation	0.2
Email Marked Spam	0.03
Approached upfront	0.01
Email Received	0.01
Resubscribed to emails	0.01
Proposed Grouping	**%**
Website	35.0
Email/Chat	34.6
Phone	28.9
Unreachable	1.4

Variable: Industry	%
Consumer Durables	77.1
NA's	21.9
Banking	0.4
Advertising	0.2
Medical	0.1
IT	0.08
Recruitment	0.08
Travel	0.07
Shipping	0.04
Construction	0.03
Export/Import	0.01
Financial Services	0.01
Food Processing	0.01
Industrial Products	0.01
Proposed Grouping	**%**
Consumer Durables	77.1
Not Listed	21.9
Not Consumer Durables	1.0

n for all variables = 7554

Note: NA means missing value.

Appendix 2. Additional Figures

Appendix 2.1. Distribution of In-flight Records

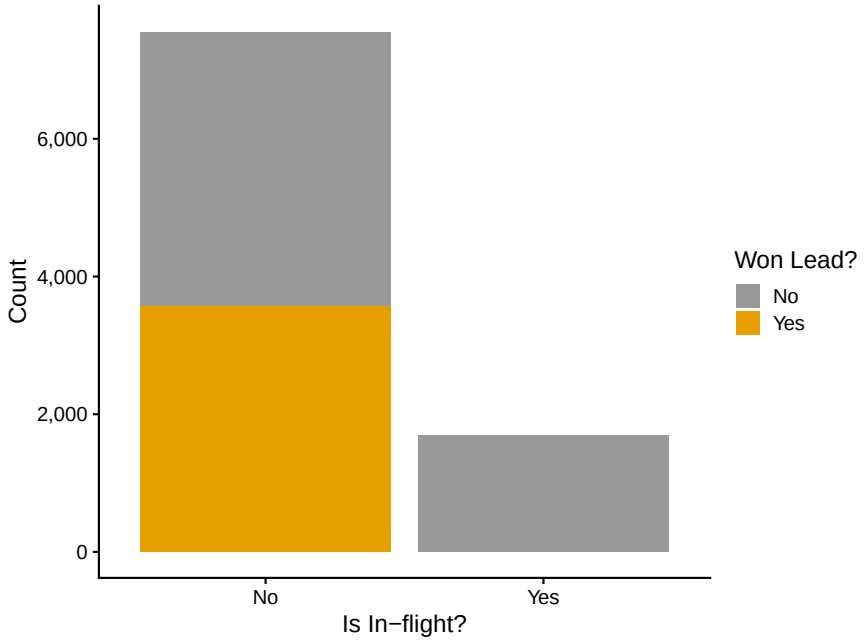

Figure 15: Records still In-flight should be removed.

Appendix 2.2. Stacked Bar Chart of Lead Stage

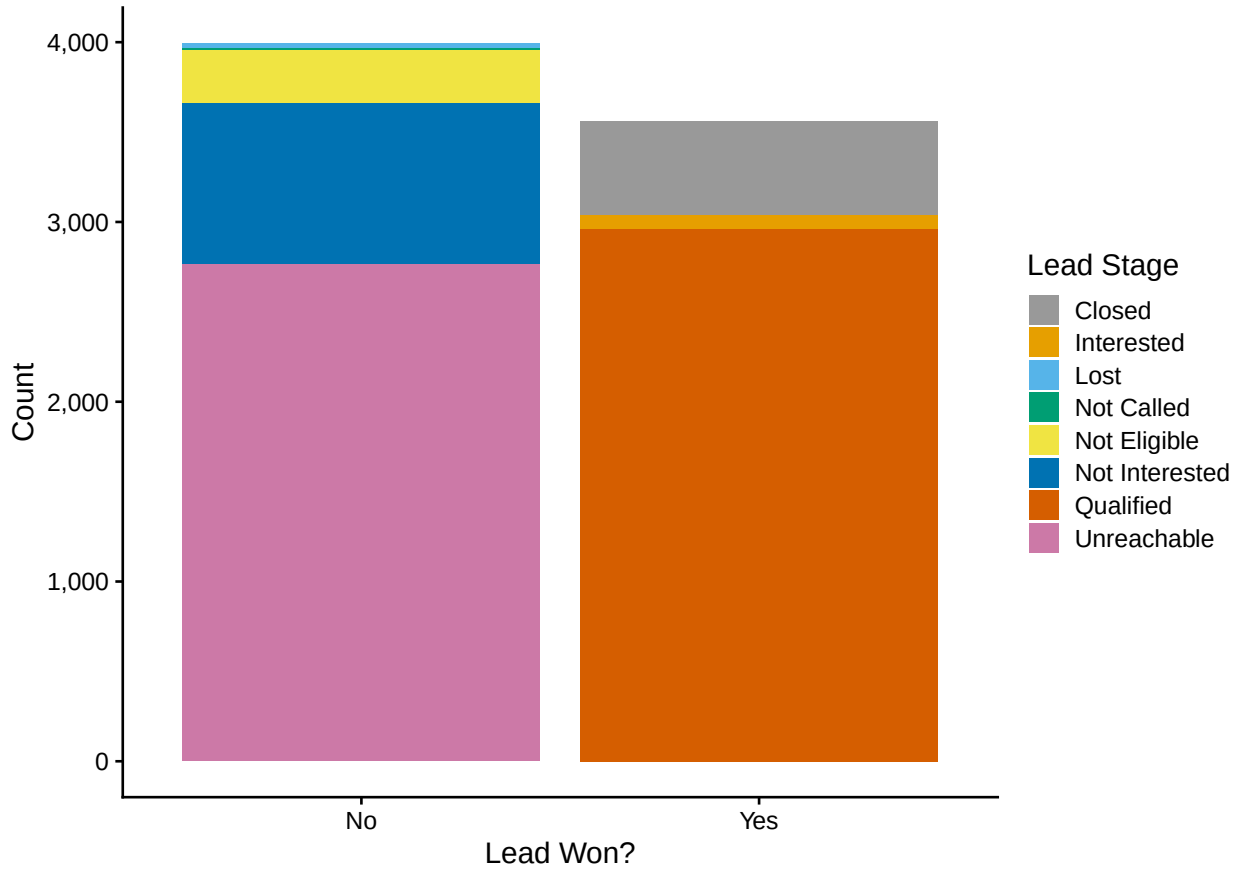

Figure 16: Stacked Bar Chart of Lead Stage.

References

Business careers with high pay : Career outlook. (n.d.). Retrieved March 05, 2021, from https://www.bls.gov/careeroutlook/2016/article/high-paying-business-careers.htm#Management.

RockBottom. (2016, August 19). Leads dataset. Retrieved March 01, 2021, from https://www.kaggle.com/rockbottom73/leads-dataset.

Interested to learn more about how Atrium can help you with predictive insights?

Contact us!

atrium.ai/contact

info@atrium.ai

4055 Valley Commons Drive
Bozeman, MT 59718

atrium