KDD Cup 2009 - Customer Relationship Prediction

by Udy Akpan, Joe Dion, Sandra Duenas, Manjari Srivastava, Jay Swinney

Summer Quarter 2015

Contents

1 Summary

This document covers the modeling process to identify three binary outcome variables from a 50,000 observation 230 variable anonymized dataset representing CRM data from Orange, a large telecom company in France. The target variables are Churn (customer attrition), Appetency (propensity to purchase) and Up-Selling (likelihood to buy more expensive goods and services). As the outcome variables are all binary, classification models including Random Forest, Logistic Regression, Naïve Bayes, support Vector Machines and K Nearest Neighbor have all been applied and compared with the goal of selecting the best model type and variable set for each of the three target variables. (I.E., the modeling process for each target has been completed independently and models and variables selected for each target will be custom selected). This is the 3rd check in and results here-in are within 1 to 4 points of the company's in-house results. The team is currently focussed on developing ensemble models to combine the results of submodels with the goal of improving the models effectiveness.

2 Introduction

Customer Relationship Management (CRM) software first became available in the 90s and has proliferated through companies large and small as a way to track interactions among companies and their customers. Whether you call Citibank, Verizon, Comcast, Carnival Cruise Lines about that cruise you are planning, Microsoft Technical support to fix a technology problem, or you are the target of an email marketing compaign, it is highly likely that your interaction will be captured in a CRM system. While the original purpose of CRM systems was to track customer interactions to closure, over time the data about customer interests, preferences and actions have become increasingly valuable and more effort is being made to extract insight to improve business decisions.

The goal of this modeling exercise is exactly that, the analysis of customer data in a CRM database with an eye towards building models to predict future customer actions. The dataset in question, the KDD Cup 2009 CRM problem is a dataset from Orange, a French Telecom company that was used as part of a KDD competition and consists of 50,000 observations with 230 variables, 190 of which are numerical and 40 of which are categorical. There are three target variables that are subject to prediction and these variables are binary, marked with either a 1 incidating the outcome occurred or exists for that observation or a -1 indicating that the outcome did not occur or exist for that observation (as part of the data transformation process, -1 values have been changed to 0 to result in binary values). The target variables of Churn, Appetency and Upselling as described below; There is no overlap among the three variables, i.e. if a customer has a 1 for churn, they will have 0s for both of the other variables.

Churn: Churn might also be thought of as attribution and in the dataset it is assumed that a 1 value indicates that a customer has stopped using the company's services. Out of the 50,000 observations only 7% have a 1 for churn.

Appetency: Represents the customers willingness to buy the service. It is assumed that a 1 value indicates a customer is likely willing to use additional services. Only 2% of observations are marked with a 1 for appetency indicating a proclivity for buying.

Upselling: Represents the likelihood of the customer to upgrade to a more profitable services. It is assumed that a 1 value indicates that a customer is likely to upgrade or be subject to an upselling marketing approach. About 7% of the observations have a 1 indicating upselling.

3 Issues with the Data

There are several significant issues with the dataset that required solutions or strategies before model development could begin.

Number of Variables: As noted above, there are 230 variables in this dataset and therefore building model to required an approach to variable selection and reduction that would produce the most effective collection of variables.

Anonymity: Several levels of anonymity have been implemented. First the variable names have been replaced by number values, i.e. Var1 to Var290. Secondly, the variable values have been replaced with seemingly non-sensical information for instance categorical vriables have been replaced with series of random characters. The random series of characters are present in more than one observation so clearly represent some type of categorical identification, but, it is not clear what that is. The actual product or service that the company is offering is also unknown.

Unknown Granularity: There are 50,000 observations in the dataset, however, it is not clear if each observation represents one customer or for instance if observations are targets of marketing campaigns where a single customer can appear more than once. For model building it is assumed that each observation represents a single customer.

Missing Variables: Many of the observations are missing values for many of the variables. Combined with the anonymity above, it is difficult to determine if data is missing for a legitimate reason, as in possibly the values represent marketing campaigns and a missing value indicates that that customer was not targeted by that campaign.

4 Missing Variable Resolution

Given that the predictors are unknown and are generically labeled, a strategy was developed for data imputation.

Numeric variables: Missing values for numeric variable was done by using zero (0) for numeric variables

Categorical variables: Missing values for categorizal variables were replaced with 'missing' and a indicator variable was added to retain visibility with a 1 indicating the value was replaced and a 0 indicating it was not.

5 The Modeling Problem

The goal is to identify the most effective set of variables and most effective model or combinations of models to predict a future customer's likelihood of churn, appetency or upselling using the availabe data. As each variable has a yes or no outcome the models used are those applicable to binary classification outcomes such as Logistic, Support Vector Machines, Random Forest, Naive Bayes and Decision Trees. Each of the target variables have been considered independently and the variable selection process was applied to each target separately.

In addition to identifying the best model for each outcome variable, as the dataset has been used in a KDD competition there is a secondary goal to exceed the result of the winning groups from that competiton using a test data set. The results of those teams below, along with information on the approach used, was considered during model development. The original competition used both a large dataset consisting of 15,000 variables and a smaller dataset with 230 variables. All comparisons will be made against the smaller dataset.

First Place: University of Melbourne (The generally satisfactory model)

Churn	Appetency	Upselling	Score
0.7570	0.8836	0.9048	0.8484

Table 1: University of Melbourne

First Runner Up: Financial Engineering Group, Inc. Japan (Stochastic Gradient Boosting)

Churn	Appetency	Upselling	Score
0.7589	0.8768	0.9074	0.8477

Table 2: Financial Engineering Group, Inc. Japan

Second Runner Up: National Taiwan University, Computer Science and Information Engineering (Fast Scoring on a Large Database using regularized maximum entropy model,categorical/numerical balanced AdaBoost and selective Naive Bayes)

Churn	Appetency	Upselling	Score
0.7558	0.8789	0.9036	0.8461

Table 3: National Taiwan University

However, the IBM Research Submission does not appear as a Winner of the Slow Track, it has the submission Score as follows

Churn	Appetency	Upselling	Score
0.7651	0.8819	0.9092	0.8521

Table 4: IBM

Evaluation: The results of the overall modeling exercise will be evaluated according to the arithmetic mean of the AUC for the three prediction tasks (churn, appetency. and up-selling). This is considered the “Score”. Larger numerical values indicate higher confidence that observations in the test set are correctly classified. The goal is to exceed the results of the in-house model which are shown below. The winning competitors from the KDD competition above only slightly beat the in-house model.

6 The Data

For the original competition, 2 data sets were used by competitiors, a large dataset consisting of 15,000 variables and a reduced data set of only 230 variables available for competitiors using personal computers rather than larger more powerful systems. As previously noted, the dataset has a number of issues. The variable names have been replace with generic names, i.e. Var1, Var2, etc, so, for each variable, there is no way to determine what the variable actually represents. There are many missing values, but imputation of missing values is even more difficult than normal as knowledge about what a variable represents could aid in selecting the imputation method.

The table below shows that there are no missing values in the target variables, values are either a 1 or a -1 (which was changed to a zero for modeling purposes) and a small number of the outcome variables have 1s, indicating the status for that observation, so, for Churn, a 1 indicates that the customer churned or left, for appetency, the 1 indicates a propensity to buy a product or service and for upselling a 1 indicates the customer has acquired additional products or services or has upgraded their products or services.

Churn	Appetency	Upselling	Score
0.7435	0.8522	0.8975	0.8311

Table 5: In House Models

	nobs	NAs	Minimum	Maximum	Q1	Q2	Mean	Median	Positive_Instances
churn	50000.00	0.00	0.00	1.00	0.00	0.00	0.07	0.00	3672.00
appetency	50000.00	0.00	0.00	1.00	0.00	0.00	0.02	0.00	890.00
upsell	50000.00	0.00	0.00	1.00	0.00	0.00	0.07	0.00	3682.00

Table 6: Response Variables

7 Exploratory Data Analysis

7.1 Imputing Missing Data

The most frequent strategy used to impute missing data was to replace missing numeric values with a 0 and create a new binary variable indicating that the variable in the original dataset was missing. For categorical variables, all classes that represent less than 1% of the total observations were grouped into an "other" category, then a separate missing class was created. The categorical variables were replaced with the word 'missing' and a new variable was added with an indicator value set to 1 to indicate that the variable was imputed or to 0 to indicate no imputation. Through this method the original variables include values for all observations and separately, a variable with missing values can be used for modeling as well in the case that the variable being present or absent is predictive.

7.2 Variable Selection

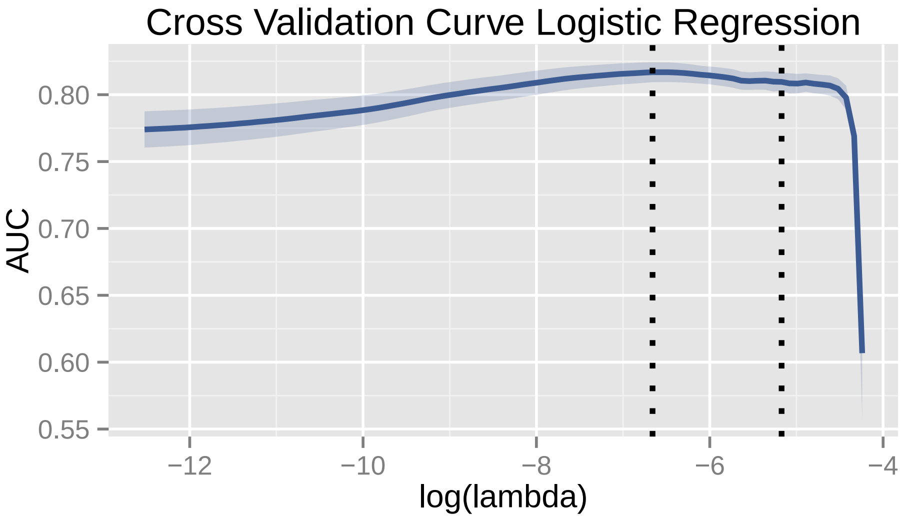
As noted previously, with 230 original variables, plus additional dummy variables to represent missing values, and variables with anonymized information, a method was needed to reduce the set of variables to be used in modeling. Logistic models with Elastic Net Penalty, Decision Trees and Random Forest were run for each of the target variables to identify the most viable variables.

7.3 Appetency

The first response variable to discuss is appetency. As defined in the task description on the KDD website, appetency is the propensity to buy a service or a product. Only 2% of observations have a positive inticator for appetency.

7.3.1 Logistic Regression with Elastic-Net Penalty

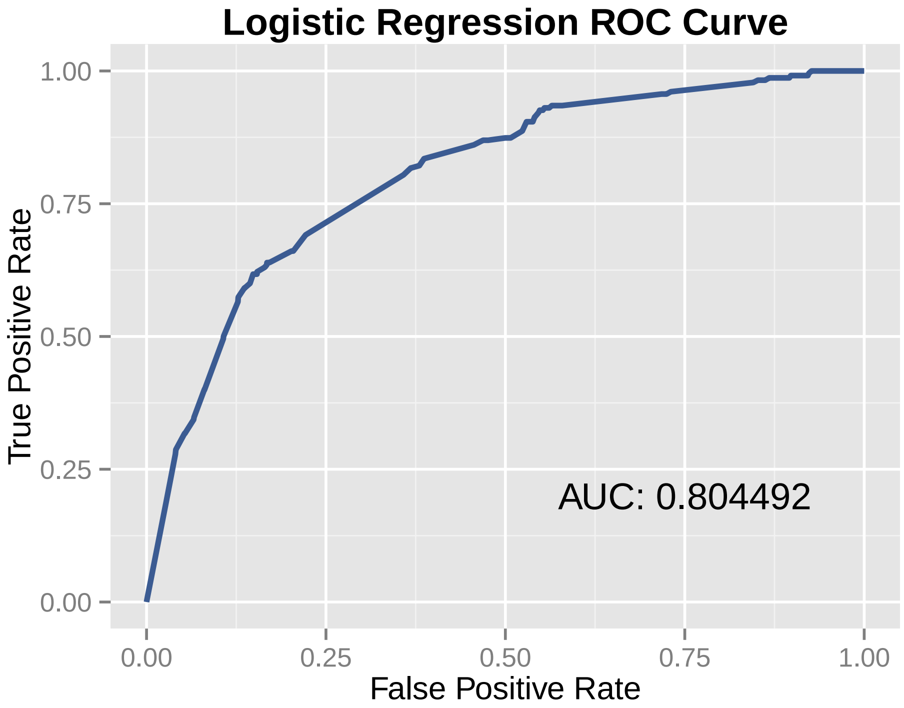
The results from the logistic regression shown below are very promising. The AUC peaks above 0.8 and does not dramatically decline until nearly all of the variables are removed from the model. This shows that a small number of variables are going to be strong indicators of appetency.



The table below indicates that with just 3 variables in the highly regularized model (right-most vertical line) Var126 and a couple of levels of dummy variable for Var218 are very indicative of appetency, meaning that predicting appetency should be relatively easy.

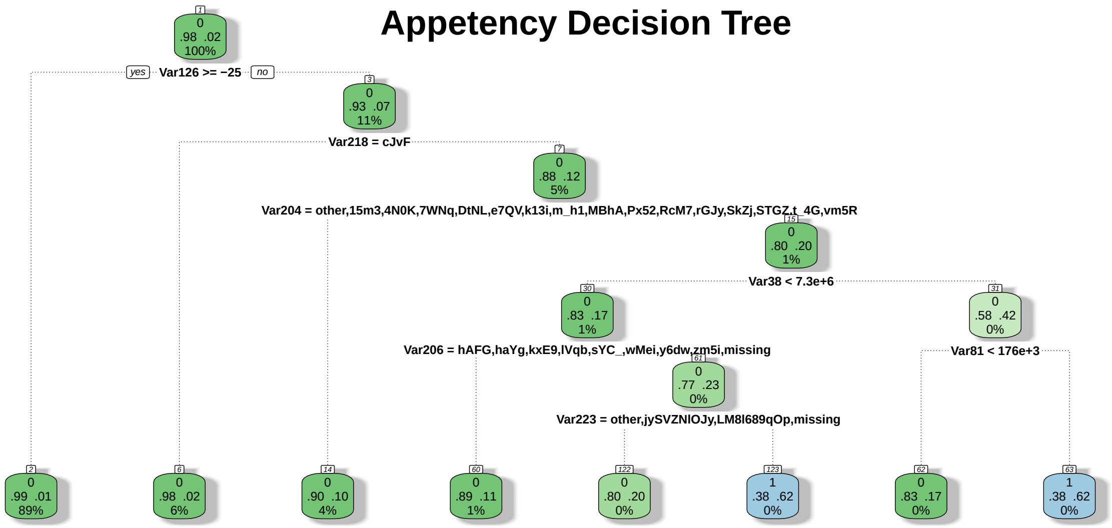
Table 7: Variables Selected by Elastic-Net

	coeficient
(Intercept)	-3.9459138
Var126	-0.5846417
Var218__dummy__cJvF	-0.7195445
Var218__dummy__UYBR	0.1148212



The ROC curve below is constructed on out of sample data, showing that the logistic regression model performs very well for appetency.

7.3.2 Decision Tree



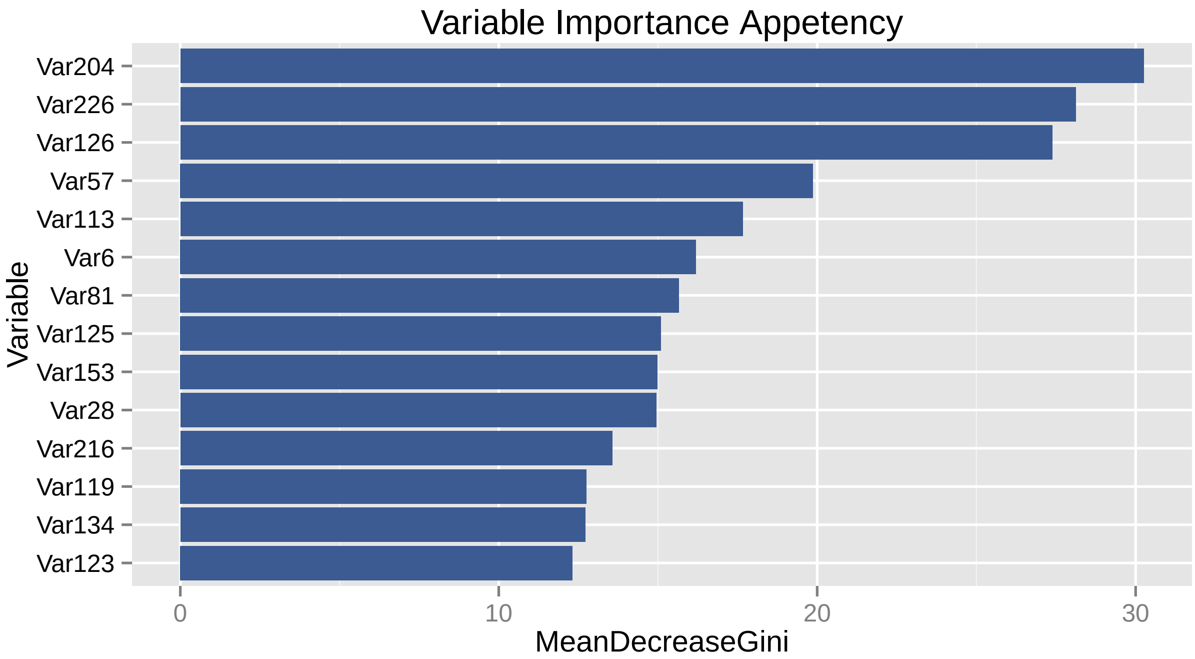
The Decision Tree classifier selected 7 variables as the most predictive, listed below in order of predictive capacity.

1. Var126
2. Var218
3. Var204
4. Var38
5. Var206
6. Var223
7. Var81

The decision tree was configured as follows: minsplit=40 to set the minimum number of observations per node, minbucket=10 to set the minimum number of total nodes, and cp=0.001 to set the cost complexity factor with a split that must decrease the overall lack of fit by a factor of 0.001.

7.3.3 Random Forest

The Variable Importance plot below for a Random Forest model identified variable 204 & 126 as two of the top three most important variables. 126 shows up in all three models and 204 also shows up in the Decision Tree.



Overall, Random Forest does not perform nearly as well as the regularized logistic regression, as the model is severely over-fit, and will need significant tuning before it reaches the level of the regularized logistic regression model.

7.3.4 K-Nearest Neighbors & Naïve Bayes

The variable selection process for appetency was based on the smallest deviance of each variable. This variable selection process resulted in 31 variables out of 230 with deviance of 186.294 or less based on the Calibration data set. The Calibration data set is a 10% random selection of observations from the original data set.

7.3.5 comparison of Variables Identified by Each Method

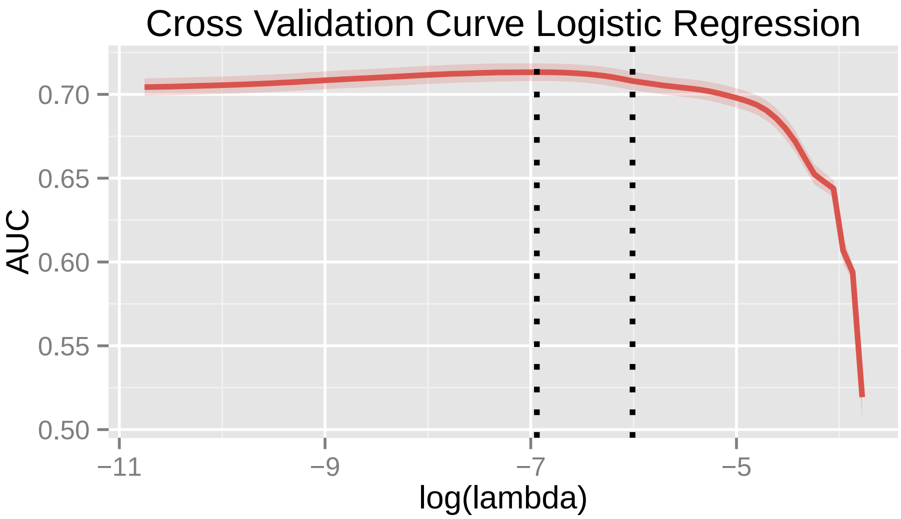
TO BE ADDED LATER

7.4 Churn

The next target variable to be evaluated is Churn, which is the likelihood that a customer will discontinue using the goods and services of the company. 7% of observations have a value indicating that the customer has churned. As with appetency, logistic regression with an elastic net penalty, a Decision Tree and Random Forest model were all applied in order to select the most useful variables.

7.4.1 Logistic Regression with Elastic-Net Penalty

Relative to appetency, the Logistic Regression model did not perform nearly as well with an AUC score in the mid 70s compared to 80 for the appetency model. Furthermore, the regularized and cross validated model selected 155 variables, many of which are the dummy variables created to indicate that values for variables were missing, in other words, whether an observation had a value for a variable or not, seems predictive, as if those variables were indicative of a attribute applicable only that those with a value. This for instance could mean that those observations received a certain marketing campaign.



For brevity, a small subset of the variables selected are displayed below with their coefficients, note that most of the variables selected are the dummy variables indicating that a value is present or absent.

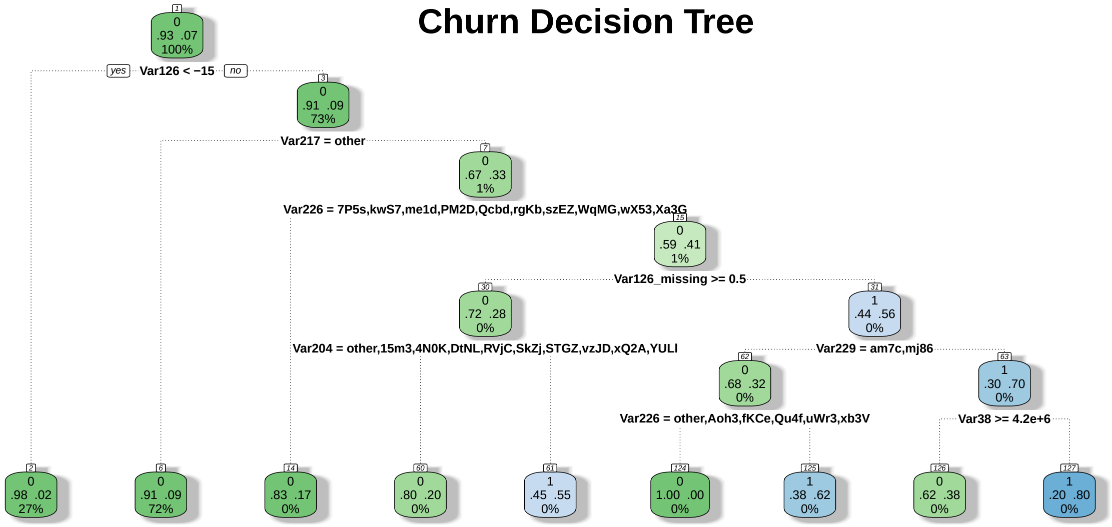
Table 8: Variables Selected by Elastic-Net

variable	coeficient
Var126	0.2553552
Var126_missing	0.3573078
Var226_dummy_FSa2	0.0643113
Var226_dummy_PM2D	0.0319844
Var226_dummy_me1d	-0.4173382
Var226_dummy_TNEC	0.1534265

variable	coeficient
Var226__dummy__uWr3	0.0840565
Var226__dummy__7P5s	-0.0186982

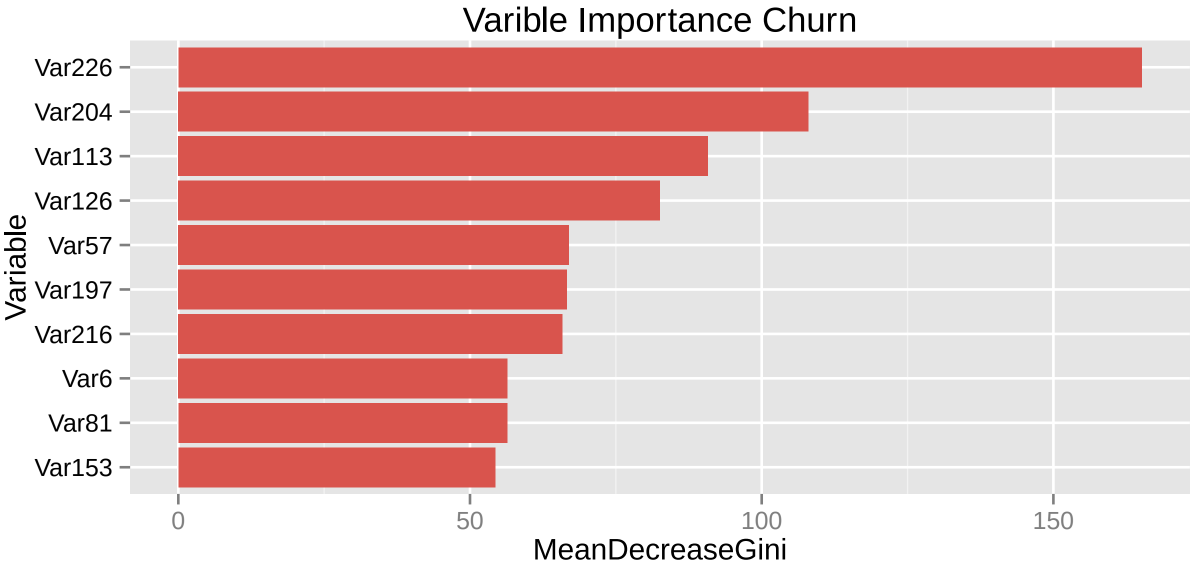
7.4.2 Decision Tree

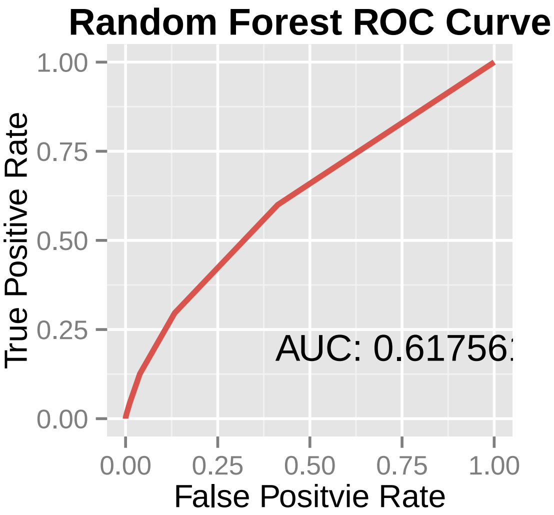
The Decision Tree model for churn is shown below. Of particular importance, variables 126 and 226 appear in more than one split and these variables were also identified as important in the logistic model above.



7.4.3 Random Forest

While the Random Forest model performs very poorly, indicating and ROC value of just over .6, the Variable Importance chart shows Var226 and Var126 as important indicating that these are likely to be highly predictive variables.





The ROC curve above is on the out of sample data and performs poorly, although the ROC curve on the in sample data, which is not displayed here, performed well. This indicates that the model is overfit and will require additional tuning. Options include changing the requirements for leaf and split sizes and trying the random forest with a subset of variables such as the ones selected by regularized logistic regression.

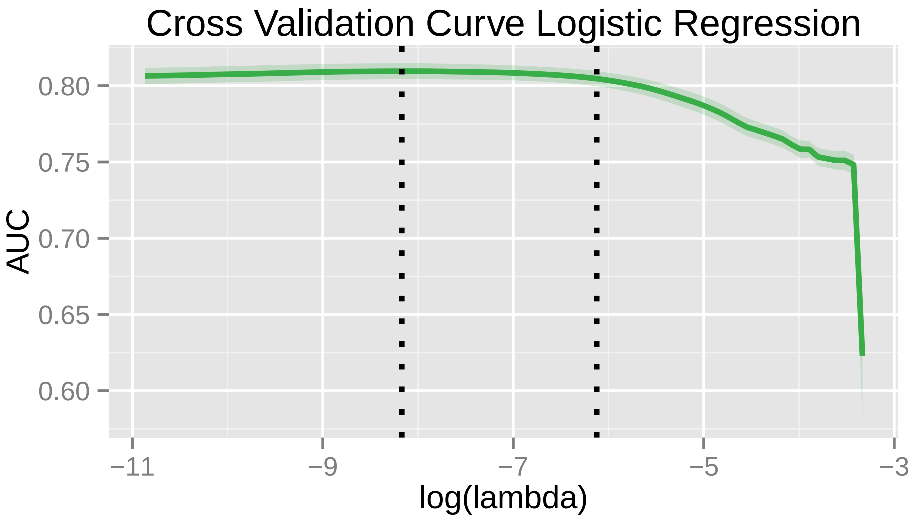
7.4.4 K-Nearest Neighbors & Naïve Bayes

The variable selection process for churn was based on the smallest deviance of each variable. This variable selection process resulted in 47 variables out of 230 with deviance of 291.862 or less based on the Calibration data set. The Calibration data set is a 10% random selection of observations from the original data set.

7.5 Up-Sell

The last response variable to be evaluated is up-sell. Up-selling indicates that the customer has purchased additional goods and services or has upgraded to a higher level of goods and services. 7% of the observations have a positive indicator for up-sell.

7.5.1 Logistic Regression with Elastic-Net Penalty



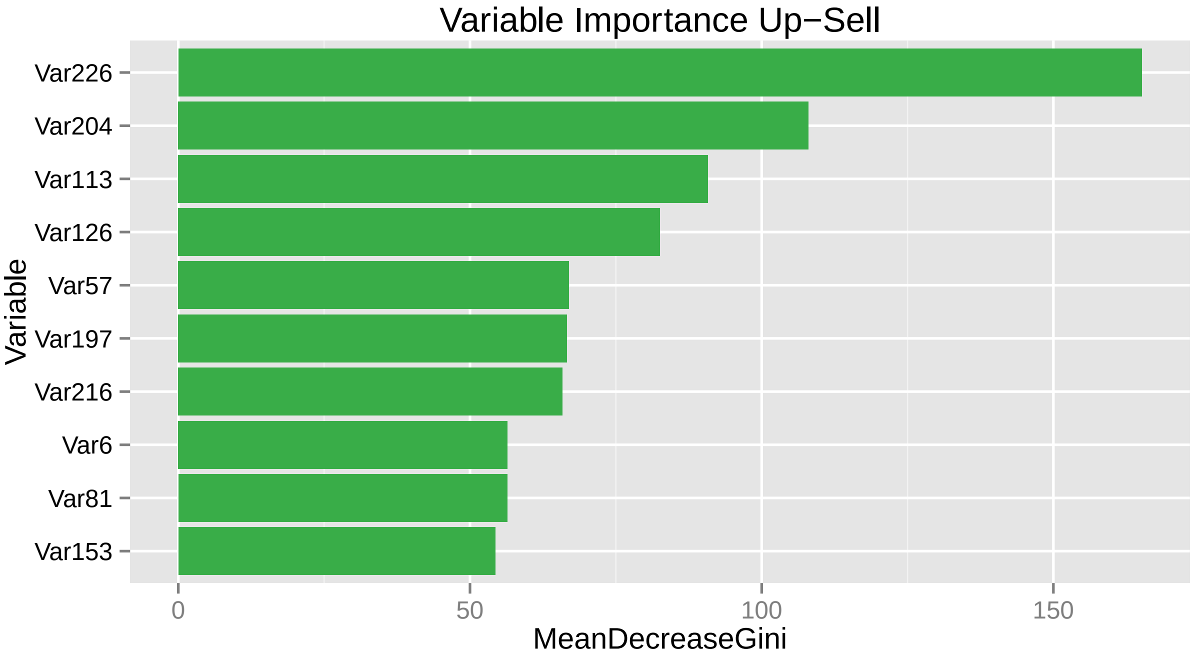
The results from the regularized logistic regression show an AUC score comparable to appetecy in the .80 range, however with 80 remaining variables in the model. Regularization does not appear to yield much performance gain.

7.5.2 Decision Tree

The Decision Tree identified Var126 and Var28 as having high importance. These variables likely have good predictive value for up-sell. Control options used for the decision tree include: minsplit set the minimum number of observations per node, minbucket to minimum number of total nodes , cp - split must decrease the overall lack of fit by a factor of 0.001 (cost complexity factor).

7.5.3 Random Forest

The Random Forest classifier selected nearly 200 predictor variables as having significant predictive value for up-sell, which does not help in reducing the number of variables for modeling. The Variable Importance plot below does includ Var126 which shows as an important variable in the Decision tree.



7.5.4 K-Nearest Neighbors & Naïve Bayes

The variable selection process for up-sell was based on the smallest deviance of each variable. This variable selection process resulted in 51 variables out of 230 with deviance of 504.483 or less based on the Calibration data set. The Calibration data set is a 10% random selection of observations from the original data set.

8 Predictive Modeling: Methods and Results

8.1 Train/Test Data

A 75/25 split was selected for training and testing and all models used the same training and testing data to allow good omparisons across the model output.

8.2 Appetency

The first series of models focussed on appetency, which is the propensity to buy a service or a product, for which only 2% of observations have a positive inticator. Multiple models where completed including Naïve Bayes, several version of Random Forest, Logistic Regression, and Support Vector Machines. A comparative ROC curve appears at the end of the commentary below and for appetency, several of the models are approaching the In-House value of the original company Orange, with one of the Random Forest producing an AUC of .84 just behind the In-House model of .85.

8.2.1 Naïve Bayes

The Naïve Bayes technique was applied in a computational EDA manner to obtain the highest AUC score for appetency. The variable selection process was based on the smallest deviance of each variable and resulted in selecting 31 of the 230 variables with deviance of 186.294 based on a calibration data set consiting of a 10% random selection of observations from the original data set.

The resulting Naïve Bayes model using the selected variables produced an overfit model with a very high AUC Score on the trining set of 0.9619 an AUC Score with the Test data of only 0.8, for a 16-point difference. Relative to the other models, Naïve Bayes performs acceptably at a .8 AUC, but, is 4 points behind the leader of .84.

8.2.2 Random Forest

The Random Forest classifier was also used for appetency with parameters set to number of trees = 50 and minimum bucket size =10 . The first model was built choosing all the 230 variables and the imputed variables. This resulted in a random forest model with over 200 variables. This model performed well on the training set at 98.6% accurate, however when applied to the testing data set it peformed poorly.

Multiple models were created using a subset of variables based on importance - top 25 and top 50 variables. These models did have a high accuracy in the training set but did not perform well on the test data set. Models were created using the sample size option (sampsize = c(10,30)) , this allows the algorithm to randomly draw between 10 and 30 from two values of appetency = 0 and 1 to grow the tree. This improved the accuracy of the model, but the results on the test data set were still not very high. Area under the curve (AUC) is used to determine goodness of fit for the model. AUC for Random forest model for Appetency in test data is 0.68, whicih is only slightly better than random selection.

8.2.2.1 Random Forest 2

A second approach to Random Forest was attempted using an alternate method. Due to the low number of positive observations for appetency, a balancing method was applied. Before fitting the Random Forest model, each observation showing a positive value for appetency in the training set was copied three times, so that there were four copies of each positive record. This oversampling of positive appetency cases was intended to make the Random Forest model predict 1 for appetency more often by changing the ratio of positive and negative cases in the data (as only 2% of the observations had a positive value for appetency). This method performed very well and is currently leading all models with an AUC score of .84.

8.2.3 Logistic Regression

Logistic regression was also completed using a complement of method. Logistic Regression with a LASSO shrinkage approach resulted in the bset model with 54 remaining variables and an AUC score of .81, making it the second best model behind the Random Forest model noted above.

8.2.3.1 Invesitigative Variable Selection

Both Decision Trees and the LASSO method were used to identified the most effective vaiables and manual variable selection was also performed to increase AUC, as described below.

8.2.3.2 Decision Tree Variable selection

Fitting a naive decision tree on the training data set produces a tree constructed using minsplit (the minimum number of observations that must exist in a node in order for a split) and minbucket (the minimum number of observations in any terminal node) is set to the values 100 and 10 respectively. The following six variables were identified as interesting with regards to appetency: Var126, Var204_dummy_RVjC, Var218_dummy_cJvF, Var25, Var38, and Var57.
A graphical analysis (not shown) of these 6 variables revealed the following:
1) Lower values of Var126(between -25 and +13) seem to be associated with high proportionate appetency

2) A higher count of appetency for observations with no values for Var204_dummy_RVjC
3) A higher count of appetency for observations with no values for Var218_dummy_cJvF
4) High counts of appetency for Var25 values below 2000
5) High counts of appetency for Var38 values below 5,000,000
6) Relatively similar counts of appetency across all values of Var57.

8.2.3.2.1 Goodness-of-Fit of Decision Tree Variables

Using the variables obtained from the decision tree variable selection step, a logistic regression model was fit on the training data set using appetency as a target. Only three variables were identified as statistically significant. An ANOVA analysis between the full model containing all 6 variables and a reduced model containing the three statistically significant variables indicated that the reduced model fits as well as the full model. The variables are Var126, Var218_dummy_cJvF and Var38. A chi-square goodness of fit test for the overall model is significant at p=0.05. The AUC score, however was below that of the model identified through LASSO.

8.2.3.3 LASSO Variable Selection

Using the LASSO (shrinkage parameter, lambda=1) a selection of 54 variables where identified when the shrinkage parameter, lambda, is at its minimum. Several variables were not statistically significant, however, an ANOVA analysis between the full model containing all 54 variables identified by the LASSO, fit better than a reduced model in which statistically insignificant variables were dropped. A chi-square goodness of fit test for the overall model is significant at p=0.05.

8.2.3.3.1 Goodness-of-Fit of LASSO variables

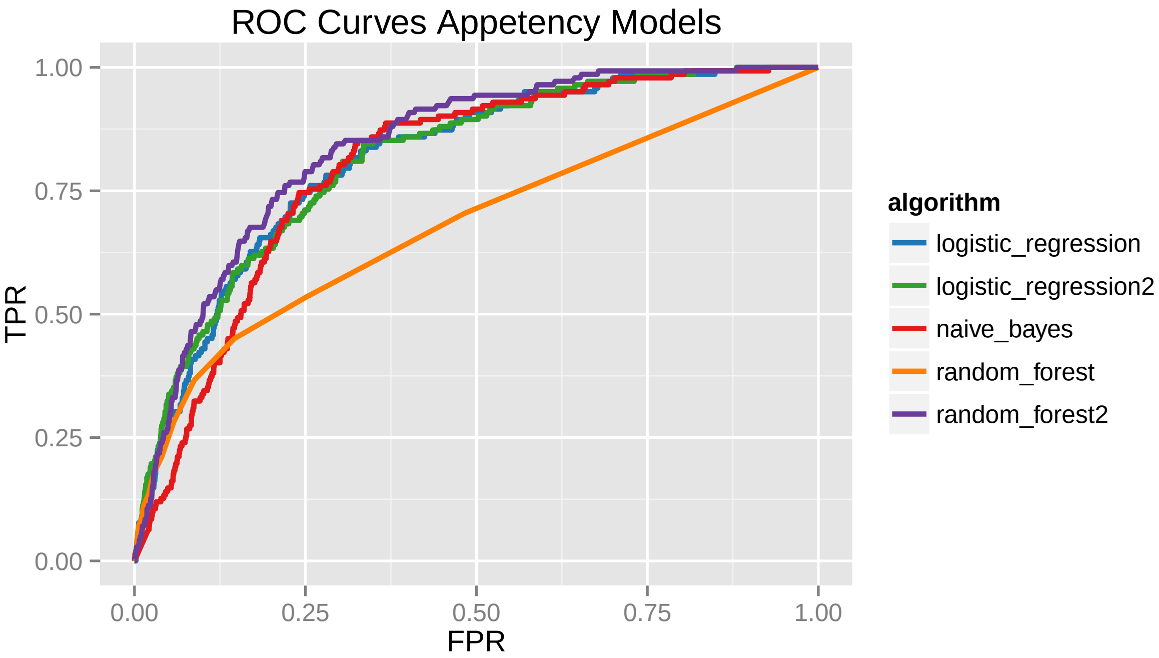
Using the variables obtained from the LASSO exploratory model, a logistic regression model on the entire data set using appetency as a target. 2 variables are inestimable and several variables are not statistically significant, however, an ANOVA analysis was completed comparing the LASSO variables to the Random Forest variables and the full LASSO variables model fit better than a reduced model in which statistically insignificant variables were dropped. We also note a chi-square goodness of fit test for the overall model is significant at p=0.05.

8.2.4 K-Nearest Neighbors

The Nearest Neighbor K technique was applied in a computational EDA manner to obtain the highest AUC score for appetency. The variable selection process was based on the smallest deviance of each variable and resulted in 31 variables out of 230 with deviance of 504.483 based on a calibration data set of a 10% random selection of observations from the original data set.

The resulting knn model used the selected variables and k = 200. The model showed over-fitting as the AUC Score with the train data was 0.9904 but the AUC Score with the test data set is only 0.6548 which while significantly above 0.50 of a random guess is lagging behind other models.

8.2.5 Model Performance



	AUC
In_House	0.85
logistic_regression	0.81
logistic_regression2	0.81
naive_bayes	0.80
random_forest	0.68
random_forest2	0.84

8.3 Churn

The next series of models focussed on Churn, which is the likelihood of a customer no longer purchasing the goods or services of the company, for which 7% of observations have a positive inticator (which in this case is a negative outcome as in the customer has churned). Multiple models where completed including Naïve Bayes, several version of Random Forest, Logistic Regression, and Support Vector Machines. A comparative ROC curve appears at the end of the commentary below and for Churn. Several of the models are approaching the In-House value of the original company Orange, with one of the Random Forest producing an AUC of .72 just behind the In-House model of .74.

8.3.1 Naïve Bayes

The Naïve Bayes technique was applied in a computational EDA manner to obtain the highest AUC score for churn. The variable selection process was based on the smallest deviance of each variable resulted in 47 variables out of 230 with deviance of 291.862 based on a calibration data set of a 10% random selection of observations from the original data set.

The resulting Naive Bayes model using the selected variables shows that the model is over fitting the data because the AUC Score with the training data is 0.9315 but the AUC Score with the test data is only 0.6622.

While the AUC for the Test is significantly above a 0.50 level of a random guess, the Naïve Bayes model for churn lags behind the other models.

8.3.2 Random Forest

The Random Forest classifier was used to build a classification model for the Churn variable on the training data set. The parameters chosen were, number of trees equal to 50 and minimum bucket size equal to 10 . The first model was built using all 230 variables plus the imputed variables, resulting in a Random Forest model with over 200 variables. This model however was not able to detect churn=1 cases in the test data set in spite of a decent accuracy level of 92.3% in test data set.

The model was then refined using the top 50 variables based on importance from the first model. This model showed higher accuracy percentage (92.4%). The model was able to detect churn=1 scenarios better than the previous random forest model. Area under the curve (AUC) is used to determine goodness of fit for the model. AUC for Random forest model for Churn in test data is 0.6883. This is a low value, so the model is better than a random pick but it is not a good model.

8.3.2.1 Random Forest 2

Because of the success of the second Random Forest attempt for detecting appetency, a similar technique was employed for detecting churn. All of the positive instances of churn were over sampled by a factor of four. This did increase the AUC for Random Forest, but only increased the AUC score from .69 to .72, this model is however outperforming the other models.

8.3.3 K-Nearest Neighbors

The Nearest Neighbor K technique was applied in a computational EDA manner to obtain the highest AUC score for churn. The variable selection process was based on the smallest deviance of each variable and resulted in 47 variables out of 230 with deviance of 504.483 based on a calibration data set of 10% random selection of observations from the original data set.
The resulting knn model used the selected variables and k = 200 but shows over-fitting as the AUC Score with the training data is 0.9801 but the AUC Score on the test data set is 0.5751. The AUC for the Test is *not* significantly above 0.50 of a random guess, and is far behind other methods.

8.3.4 Logistic Regression

Following the same process as Logistic Regression for appetency, several approaches were used in order to select the most impactful variables for the logistic model including a Decision Tree and several variations of the LASSO method.

8.3.4.1 Decision Tree Variable Selection

Logistic Regression modeling started with variables that were selected by a decision tree. 12 variables were identified as inestimable and were dropped from further consideration. there were also a number of variables are not statistically significant which were also dropped.

9 variables remained which were used to fit the logistic regression model: Var126, Var217_dummy_missing, Var211_dummy_L84s, Var73, Var126_missing, Var229_dummy_missing, Var113, Var22_missing, and Var65. The chi-square goodness of fit test produced a p-value=0.2439 and the Logistic Model using the Decision Tree selected variables was not found to have significant predictive capacity.

8.3.4.2 LASSO Variable Selection

15 variables were identified at log (lambda), one standard error from the minimum using the LASSO method and were the following: Var7, Var73, Var113, Var126, Var22_missing, Var28_missing, Var126_missing, Var205_dummy_sJzTlal, Var206_dummy_IYzP, Var210_dummy_g5HH, Var212_dummy_NhsEn4L, Var217_dummy_other, Var218_dummy_cJvF, Var218_dummy_missing, and Var229_dummy_missing.

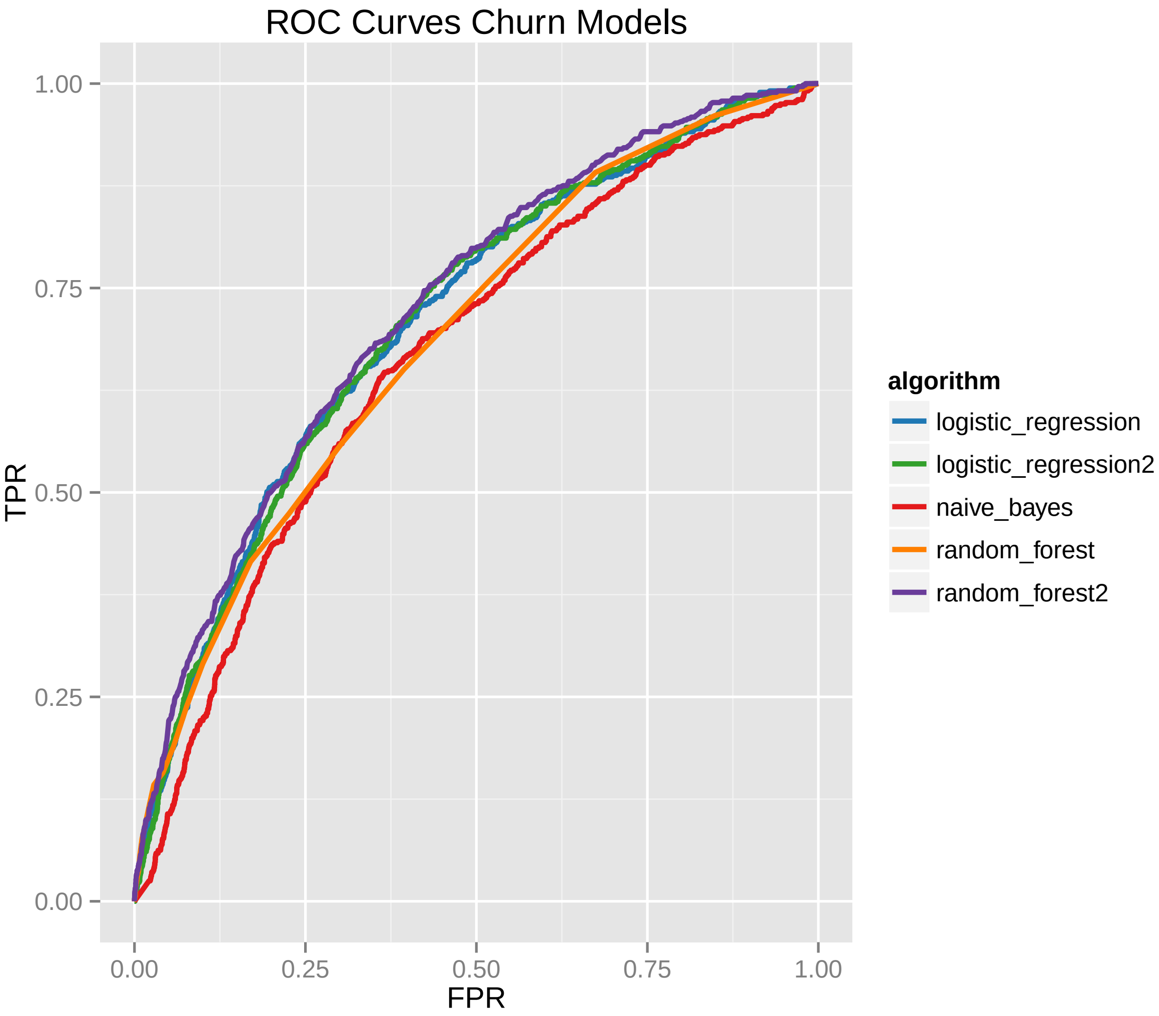
8.3.4.3 GOF on LASSO variables

Using the variables obtained from our LASSO exploratory model, a logistic regression model was fit for churnt. Several variables were either NA or insignificant and therefore elminated from further consideration. This resulted in the following 10 variables that were used in fitting logistic model: Var7, Var73, Var113, Var126, Var205_dummy_sJzTlal, Var210_dummy_g5HH, Var212_dummy_NhsEn4L, Var217_dummy_other, Var218_dummy_cJvF, and Var229_dummy_missing. The chi-square goodness of fit test produced a p-value=0.6240 and this approach was also discontinued.

8.3.4.4 Simple LASSO

Finally, a simple logistic regression model with LASSO shrinkage was fit including all of the variables. The results applied to the test data set show produced an AUC score of .71 which is within one point of the best model identified and within among the best models identified for churn. In the ROC curve this method is identified as logistic regression 2.

8.3.5 Model Performance



	AUC
In_House	0.74
logistic_regression	0.71
logistic_regression2	0.71
naive_bayes	0.67
random_forest	0.69
random_forest2	0.72

8.4 Up-Sell

The final series of models focussed on upselling, the propensity of the customer to purchase more expensive goods and services of the company, for which 7% of observations have a positive inticator. Multiple models where completed including Naïve Bayes, several version of Random Forest, Logistic Regression, and Support Vector Machines. A comparative ROC curve appears at the end of the commentary below and for Up-sell. AUC model results for Up-sell lag the results of other models with 4 points between the best model at .86 and the In-House model of .90.

8.4.1 Naïve Bayes

The Naïve Bayes technique was applied in a computational EDA manner to obtain the highest AUC score for up-sell. The variable selection process was based on the smallest deviance of each variable, resulting in 51 variables out of 230 with deviance of 504.483 based on a calibration of a 10% random selection of observations from the original data set.

The resulting Naive Bayes model using the selected variables shows over-fitting as the AUC Score with the training data is 0.9177 but the AUC Score with the test data was only 0.7515. The AUC for the test data is significantly above 0.50 of a random guess, however the model lags the results of other models.

8.4.2 Random Forest

The Random Forest classifier was also used to build a classification model for up-sell on the training data set. The parameters chosen were number of trees = 50 and minimum bucket size = 10 . The first model was built using all 230 original variables plus the imputed variables and resulted a Random Forest model with over 200 variables. While the model performed well on the training data set, it did not do well on the test data set. A second model was built using the top 50 variables based on the variable importance from the first model and the method of balanced sampling, duplicating observations with a positive value for upselling as previously described was also used. This model showed higher accuracy in the test data set. The model was able to detect up-sell=1 scenarios better than the first model with an AUC Score on the test data of 0.86 and this result is displayed in the ROC curve as random_forest2.

8.4.3 Logistic Regression

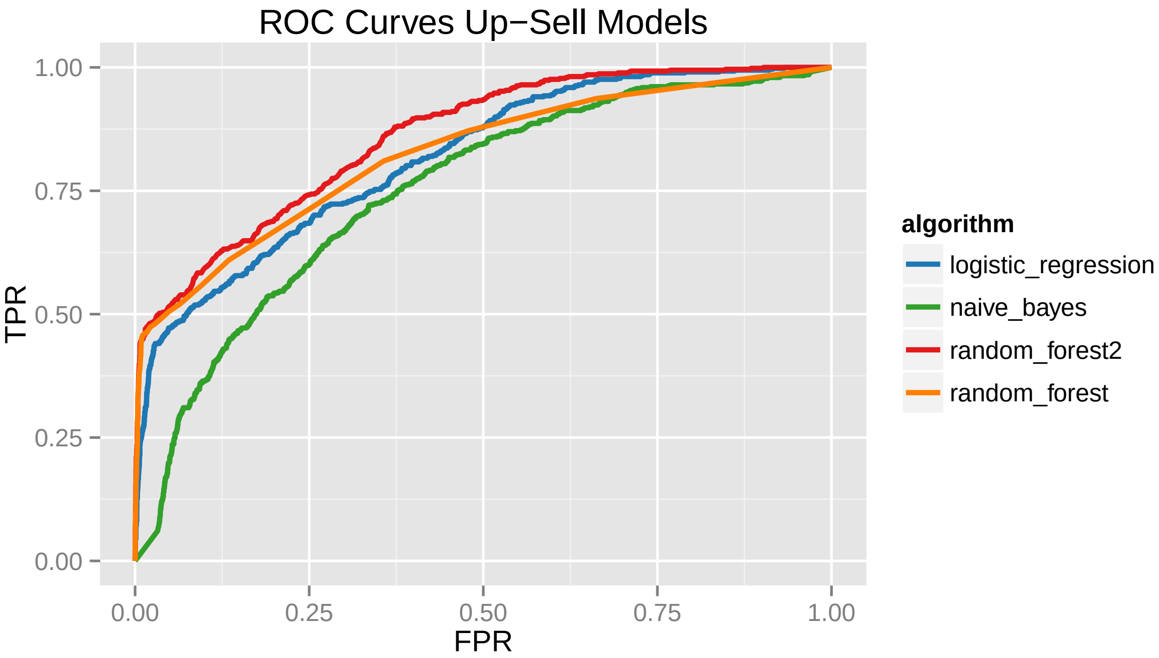
A logistic regression model was fitted for with a LASSO shrinkage parameter. Several attempts at feature engineering were made, three interaction variables were added based on the results of the decision tree discussed in the EDA portion (variable 126 and 28, variable 28 and 153 and variable 125 and 81). Also a squared version of every numeric variable was added to the data. This created a very large data set and the model had to be trained over a period of several hours. The results showed an improvement over some other algorithms, but overall results failed to match other models and further work was halted.

8.4.4 K-Nearest Neighbors

The Nearest Neighbor K technique was applied in a computational EDA manner to obtain the highest AUC score for up-sell. The variable selection process was based on the smallest deviance of each variable and resulted in the selection of 51 variables out of 230 with deviance of 504.483 based on a calibration data set of a 10% random selection of observations from the original data set.

The resulting knn model used the selected variables and k = 200, shows over fitting as the AUC Score with the training data is 0.9878 but the AUC Score with the test data is 0.7021. The AUC for the testing data is significantly above 0.50 of a random guess, but the results lag other models.

8.4.5 Model Performance



	AUC
In_House	0.90
logistic_regression	0.82
naive_bayes	0.75
random_forest2	0.86
random_forest	0.82

9 Comparison of Results

TO BE ADDED IN FINAL ROUND

10 Conclusions

TO BE ADDED IN FINAL ROUND

11 Appendix

For the interested reader, all of the code used to create the models can be found at:
https://github.com/jayswinney/454-kdd2009

12 Next Steps

A number of items remain to be added for the final submission the most significant of which are ensemble models which are being developed to combine the results of the best performing models to date.

1. Comparitive variable tables for variable selection and modeling
2. Ensemble Models
3. Comparison of Results
4. Conclusion