Andrew Porter

Exam 2

**A.**

A brief data exploration:

There are 12,733 customers who did not churn, while there were 3,596 customers who did churn.

There are two variables “setprcm” and “income” that are missing values.

See attached PDF for correlation matrix of numerical values. There are a few features that have high correlation values for instance blckvce (mean number of dropped) and dropblk (mean number of blocked) and mou (mean monthly minutes of use) and mourec (mean unrounded mou received voice calls)

Distribution analysis

Kurtosis:

Overage, Roam, threeway, refer all have large values indicating sharper peaks and flatter tails on distribution.

Skewness:

Almost all have a positive skew meaning that right side of median in longer, more larger outliers.

Model building

Based on experience, the first model to be analyzed is the **random forest** algorithm with 500 trees and 4 variables on the balanced dataset and all features. This achieved an AUC of .6641 on the validation set and an AUC of .6314 on the testing set.

**Running a level of variable selection:**

PCA with all dataset. Ignoring categorical and imputing median value for all other missing values. Doesn’t really provide much insight, we would need PC12 (12 variables to explain 70% of the data) within those 12 pseudo variables there aren’t any that would drop off based on having a low coefficient.

Executing a **logistic** model with the entire data set achieved an AUC of .6579 which is the second largest AUC. Tried to improve the model with parsing the data based on statistically significant variables only (“\*” from rattle) and achieved about the same AUC, nothing more.

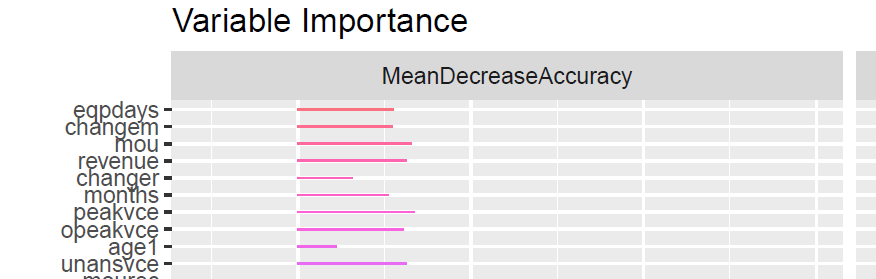
Running a **random forest** model to understand variable importance:

Based on output from variable importance the following variables are most important. Keeping the top 10 variables did not help improve the AUC. Best model was a neural net with 3 hidden layer nodes that achieved a .6466 AUC on the validation set.

**SVM radial** achieved .5467 and **SVM polynomial** degree = 2with unbalanced dataset achieved AUC of .5284, both with a value of C = 1

**SVM radial** C = 1 with balanced dataset achieved AUC of .5948.

\*note: did not run and additional SVM models because rattle’s performance is significantly slow\*



Executing the **random forest** model with entire dataset AUC achieved was .6234, with variable selection from variable importance below model achieved AUC of .5680.

**Boosted ensemble** with variable selection from random forest achieved .602

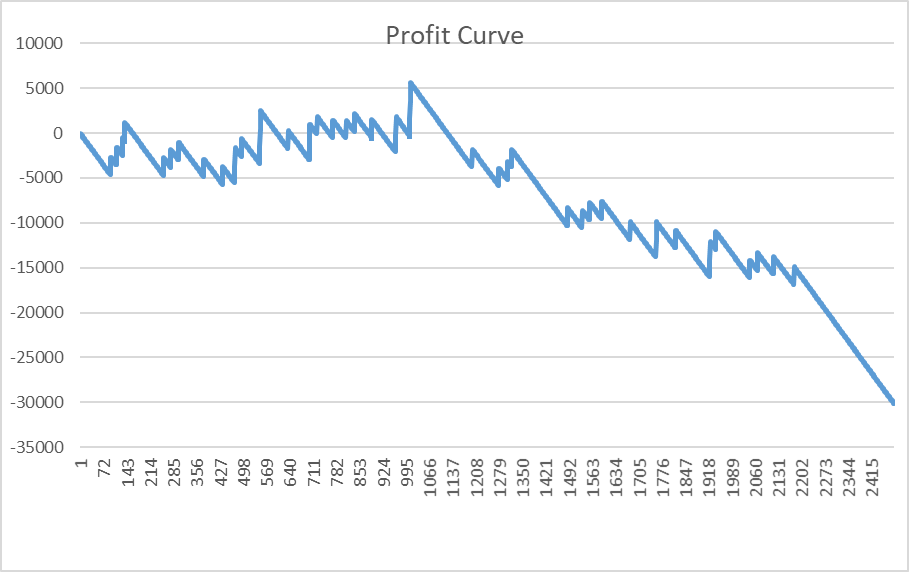
**Boosted ensemble** with all variables and balanced data set achieved .6366 on validation set.

**Outcome**

Based on all models executed, the best performing model on the validation set, based on AUC is the **random forest** model with a balanced dataset having 500 trees and 4 variables. This model achieved an AUC of .6314 of the testing set.

**B.**

To maximize the profit based on the profit curve (see associated excel with detail) Rozmova should target 1,008 customers and the profit that will return is $5600. A random forest was chosen and executed on the balanced data set to build the model and then evaluated on the test set to provide the associated spreadsheet for the profit curve.



**C.**

If Rozmova only has $10,000 for its intervention discount campaign and spent $50 on the interventions that would allow Rozmova to select the top 200 customers (with the greatest probability of churning). Based on the profit curve and examining how many true churners exist in the top 200, the company would only convert 4 customers. So this might not be the best approach.

Since there is a maximum profit of $5,600 dollars based on the profit curve. If the company only has $10,000 to spend and using the model of the 200 most likely to churn customers only retaining 4 that would have left and yielding $2,000 per customer, so $8,000 total from retaining those four customers and spending another $50 on all 200 the company would lose $2,000. Since the churn rate is so low (55 true positives in the testing set), the company would have to evaluate what their true loss would be if they lost these customers. So one could make an argue that you simply don’t campaign.

Another approach would be to convince the “decision maker” or your manager that your model has predicted a maximum profit on $5,600 dollars if you could campaign to the 1,008 customers. Which means you would be asking for $50,400 and that you would make the remaining money back by the monthly revenue from the 28 customers sticking around returning $56,000.