Project 1 – Who is likely to churn?

# Churn

You are competing for a job as data scientist working for a mobile phone company like Verizon, Mint Mobile or ATT. As part of the interview process, you have been tasked with identify which customers are likely to “churn”. Customer churn, also known as customer attrition, occurs when customers stop doing business with a company or stop using a company’s services. The company believes that churn is influenced by the network speed 5G or 4G LTE and the phone model. They also think that customers who opt for paperless billing are less likely to churn. The company uses an annual valuation of $1200 per customer so if a customer churns (i.e. switches service) they are out a considerable sum of money.

Your task is to build a **logistic regression** and a **decision tree** machine learning model to identify likely churners, so that the company can incentivize the churners to stay with the company. To do this they are offering $50 voucher to everyone your model identifies as churn, for every correct classification (churners identified) they expect to make $500 by spending $50 and for every false positive they’ll lose $50 and for every false negative they’ll lose $150.

You have been provided two datasets:

* Churn\_training.csv – use this one to train and evaluate your model.
* Churn\_holdout.csv – use this one to “score” and post to Kaggle.

The job is yours if you can beat the top candidate’s model which has accuracy of 96%, an Area Under the Curve (AUC) of 0.937, and a TPR of 70% at a 2% FPR on the churn holdout data set.

Can you beat it!

## Deliverables for Grading

For this project, you will complete and submit the following.

1. **Executive Summary &** **Model Report** - The report will detail the data munging/shaping, data understanding, data preparation and modeling phases of this project. In the report template (*project\_template.docx*) provided to you, you will basically perform the steps outlined below, document your results and answer any questions that correspond to the Required Tasks (instructions) below.
2. **Notebook** – I’ll want your notebook used to produce results in the report. Your code should be appropriately commented so I can tell what is going on – ideally, I should be able to repeat your analysis.
3. **Predictions file** – We want you to apply your model to the Churn\_holdout.csv data set this is what you’d give to your new potential employer. You will submit this file to Kaggle to see if you beat the benchmark of 0.937.

## Required Tasks

You will need to write a Model Report which will contain two major sections an Executive Summary and Detailed Analysis. And you will need to submit your predictions to Kaggle

## Executive Summary (20 points)

Here is a little truth, most executives can’t or won’t read typical analysis! Your challenge is to concisely present your findings and results without writing a book. A good executive summary will leave the reader with **actionable** takeaways that they can remember and regurgitate at the next meeting. Your executive summary should be just that a ***summary***. What problem are you challenged with? What were 3 or 4 key findings (things you found interesting that influenced the model). What was result of your model, and 2-3 recommendations that you’d make.

* State the problem
* 3-4 Key findings
* Model Performance & Interpretation of it.
* 2-3 actionable Recommendations

Helpful hint: do not attempt to draft an executive summary until after you’ve done the analysis and written the modeling report. The executive summary should be the last thing you should do.

## Detailed Analysis (25 points)

This is the meat of your report, you should follow the steps provided in the report\_template.docx document for details and steps. Your Detailed Analysis is not about getting you to write a long report really it is about communicating your understanding and findings.

Every customer/company, I have ever engaged with or worked for, asks for these type of reports – the detailed analysis documents your model development process, your assumptions, features included/excluded, feature engineering steps, modeling method, hyperparameter choices, model evaluation, alternatives attempted and final model strengths and weaknesses.

**Grading:**

This project is worth a total of 50 points

* Executive Summary is worth 20 points
* Detailed Analysis is worth 25 points
* Kaggle AUC above 0.937 is worth 5 points

Your grade will be based on the following:

* Rigorous application and documentation of the machine learning process.
* Clear, concise, and accurate discussion of the results.
* Clear explanation of how the results can be used to address the business problem at hand
* Beating the benchmark Random Forest model AUC 0.937.

### Kaggle Instructions

Use your **WFU email**, register on Kaggle, the URL below is a link to the competition.

### <https://www.kaggle.com/t/6f7f7b1a7126416091fc4d23834f1d62>

### Some expectations:

* Make sure any chart you create has:
  + Title, x & y axis labels, and a description of why someone would care to look at it.
* Make sure any table you create has
  + A description of how to interpret it.
* Make sure you always evaluate two or more models and compare them.
* Make sure you spend time understanding what your code does.