Customer Retention

By Being Proactive

A Data Science Project By Aparna Shastry

Overview

- The Business Problem
- Data / Data Wrangling
- Data Stories
- Predictive Models
- Solutions Proposed
- Scope for Further Work

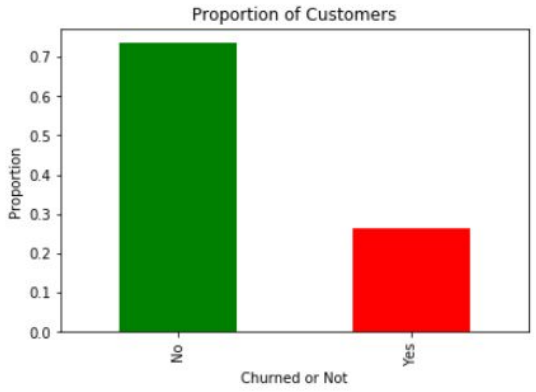
The Business Problem

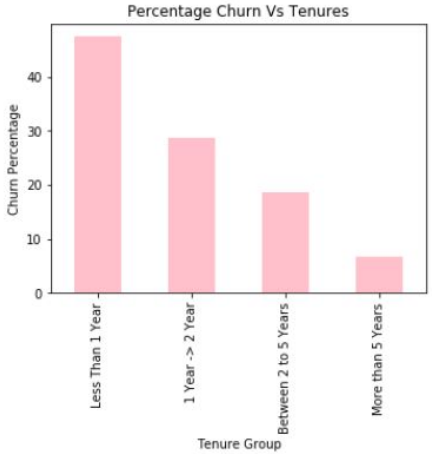
- A Telecom company observes customer “Churn”
- Estimated loss of income: $140k per month
- Attraction of new ones is more expensive than retention of current ones
- Expectations from this project: Given historical data on loyal and churn customers,
  - Understand relation between churn and certain factors
  - Provide a predictive model that ranks the customers
  - Make recommendations to the business to minimize the revenue loss.

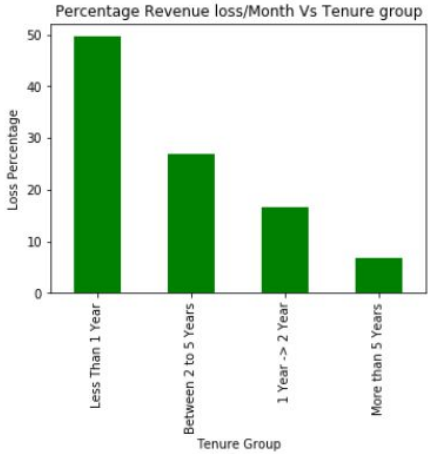
Data / Data Wrangling

- Dataset link is given in Ref [1] in the last slide.
- 7043 records, 21 columns
- The target variable: Churn : { “Yes”, “No”} entries.
- 19 Predictor Variables, Service specific, Person specific, Money specific
- Cleaning of Total Charges:
  - 11 missing entries, all of them are in rows with Churn “No”
  - tenure values are 0 for them
  - No other information is given => just registered
  - Set Total Charges to 0
  - Converted all other entries from string to float
- Observed : Total Charges is almost = tenure * Monthly Charges,
  - Total Charges is redundant
- Monthly Charges is linear combination of services => Correlated with all others

Business Problem visualization

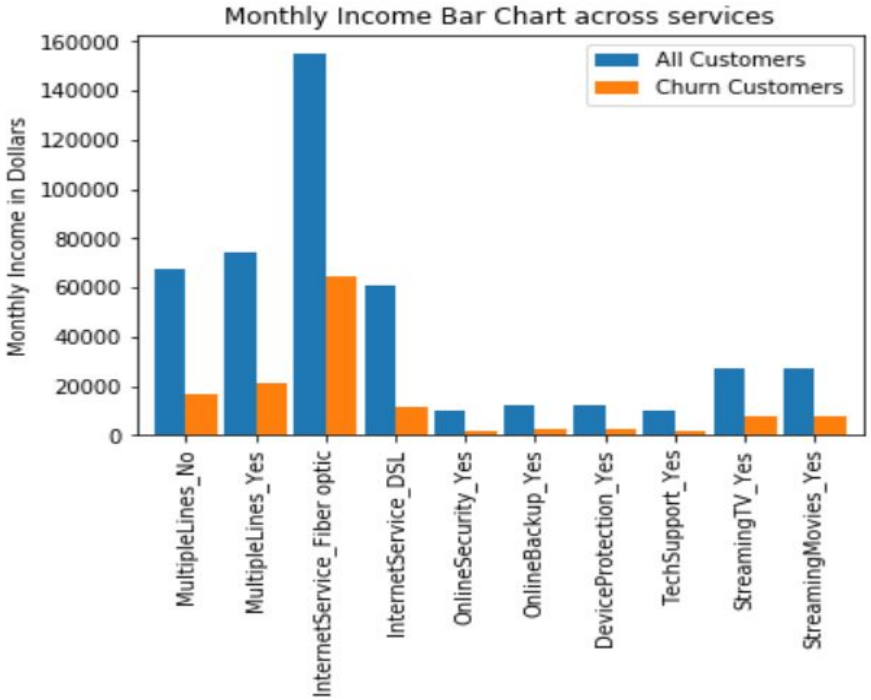






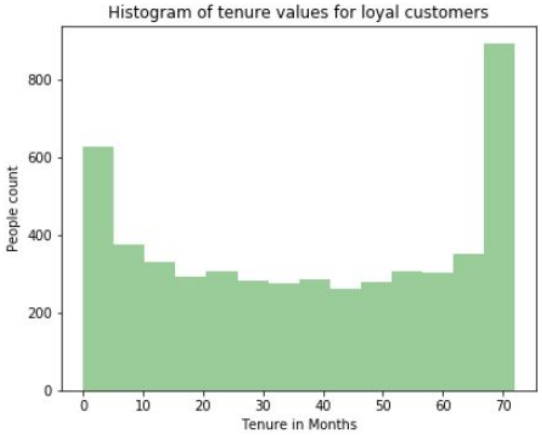
- 26.54% Churn rate overall
- $139K loss per month / 1.67M loss per year, about 30% of the total income
- Less than a year tenure category has highest churn and result in highest revenue loss
- 2 to 5 year tenure category results in highest revenue loss

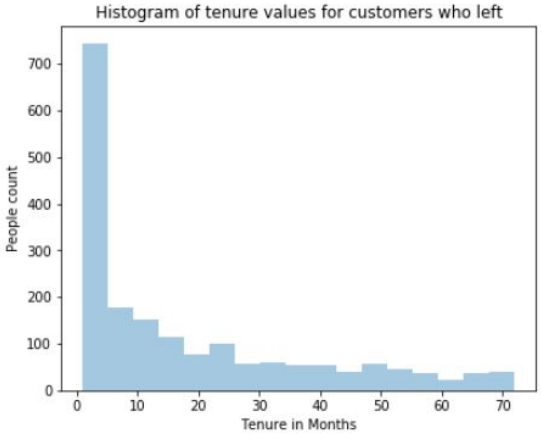
Data Stories: Monthly Charges Split by Services



- The height of the bar = count of people subscribed * Charge as per rate table of previous slide
- Blue bars are from the entire 7043 rows
- Orange bars are from the Churn group
- The importance of Fiber Optic Service on the Monthly Income is very clear!
- MulipleLines_No essentially means Phone Line Single

Data Stories : Tenure in Months





- Fig 1: Drop from bin 70 to bin 60 => Huge churn happened about 5.5 years ago.
- Fig 1: Between 10 to 60 months, not much variation => Once they cross a year, they remain loyal
- Fig 2: High count in first bin, drastic drop=> Most in churn group < 5 months tenure

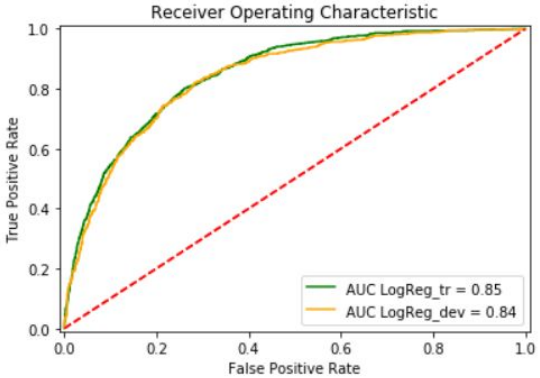
Predictive Modeling

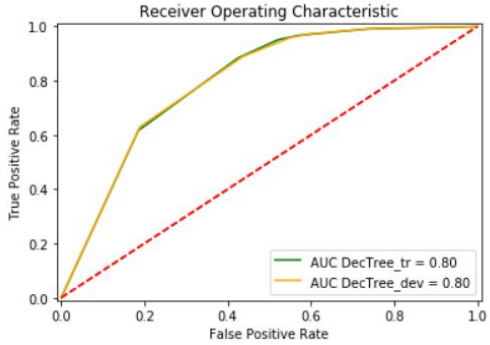
- Acceptance Criteria:
  - Train and Test Accuracy approx equal or not.
  - “Recall” on the Churn class to be higher priority than overall accuracy
- Data Prep/Feature Selection :
  - Drop numerical variable Total Charges, and Monthly Charges
  - Drop the column Phone Service, as it is a subset of Multiple Lines
- Data split to Train : Test ratio 60:40 , stratified with target
- Hyperparameter tuning done using 5-fold cross validation(CV) method on Training data
- Score function used in CV is “Recall”

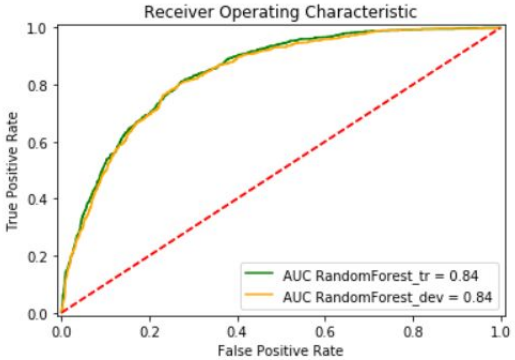
Comparison of Models Tried

Model Name	Hyperparameters Selected	Sensitivity (Recall) on the Churn group, Overall Accuracy
Logistic Regression Selected	**C = 0.1, class_Weight = balanced**	**0.80,0.75**
Decision Tree	max_depth=3,min_samples_leaf=1,class_Weight = balanced	0.89,0.65
Random Forests	max_depth=3,min_samples_leaf=1,class_Weight = balanced,n_estimators=100	0.83,0.72

Comparison of Results, ROC Plots

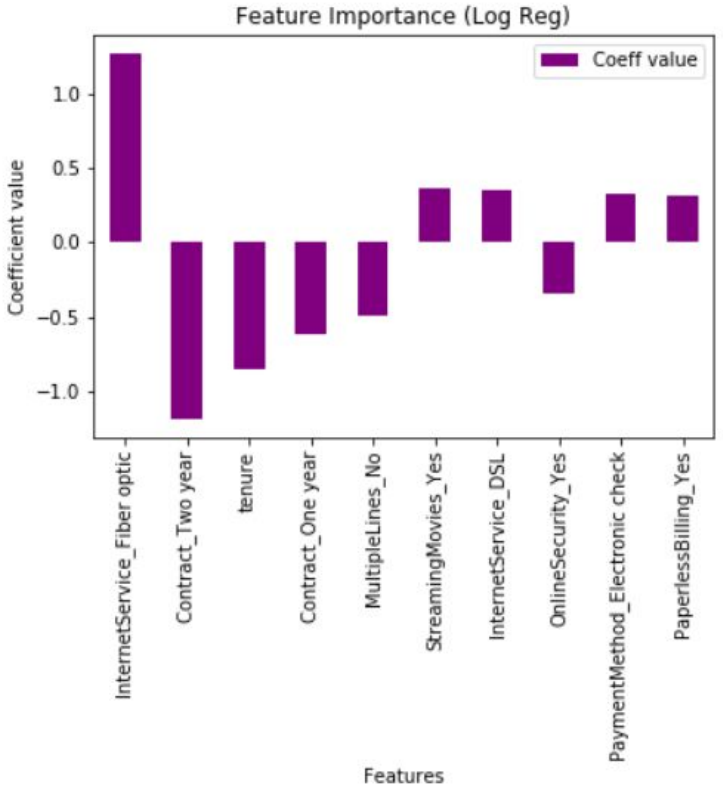






- Which one to choose: I would choose Logistic Regression because,
  - It is faster and less complex
  - More interpretable for this problem, gives coefficients with signs
  - It gives better overall accuracy than Random Forest, and almost same recall on Churn class

Feature Importance Based on Logistic Regression



Top 10 Features with their influence shown

- Interpretation:
  - Features with negative coefficients are favorable, and positive not favorable

Solutions Proposed

- A predictive model is given that ranks customers based on their probability of churn and the revenue that they bring - (not just the probability)
- Use this model to prioritize whose concerns to be addressed first. Sometimes it might be case by case basis.
- Take the following actions immediately:
  - Try striking a longer contract with new customers: two year or one year in that order of preference.
  - Leverage the time to improve the quality of services, of the high cost ones like Fiber optic.
  - Improve on the Technical support on all services like streaming, phone connection and internet.
  - Be up-to-date with current technology.
- Next: It will be helpful to understand why churn started 5.5 years ago. Give more historical data to the data scientist for analysis.
- The income loss prevented using the given dataset: $114,422 / month or 1.38M per year if all of them change their mind.

References/Links

1. IBM page has many, out of that this Telecom dataset
2. Introduction to Statistical Learning by Gareth James et. al
3. Predicting Customer Churn using R by Susan Li
4. Techniques to Handle Imbalance by Jason Brownlee
5. My Code