Customer Retention

By Allocating Resources Wisely

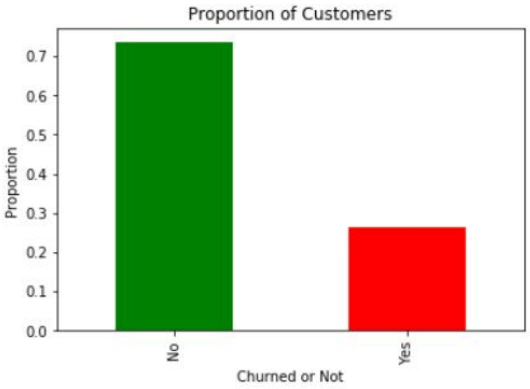
Overview

- The Business Problem
- Data Sets
- Data Exploration
- Predictive Models
- Solutions Proposed
- Scope for Further Work

The Business Problem

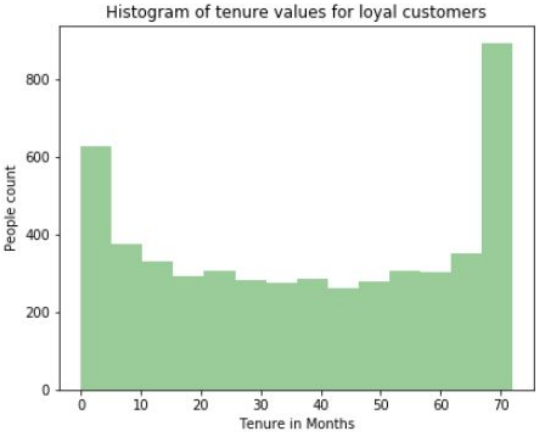
- A Telecom company observes customer stop subscription to their services: called “Churn”
- Attraction of new ones is more expensive than retention of current ones
- Expectations from this project:
  - Given customer information, predict if a customer is likely to switch
  - Figure out the top 10 influential factors of customer churn
  - Make recommendations to the company to make certain changes, allocate resources wisely for increased customer retention.

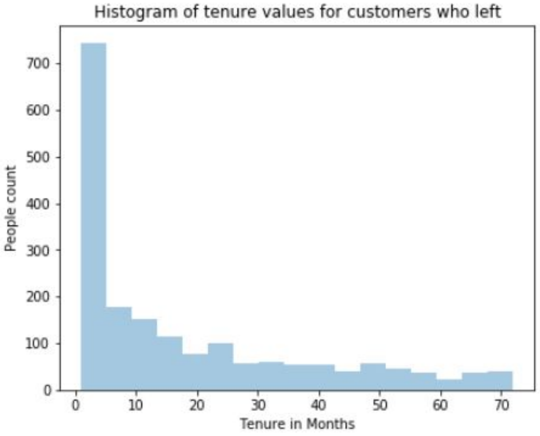
Data Set



- Dataset link is given in Ref [1] in the last slide.
- 7043 records, 21 columns
- The target variable: Churn : { “Yes”, “No”} entries.
- Churn rate (Churn = “Yes”) is 26.54% of 7043 records
- One column has unique customer ID
- 19 Predictor Variables.
  - 3 of them, tenure, monthly charges and total charges numerical.
  - Rest are categorical with 2,3 or 4 categories.
- Only Total Charges has 11 missing entries. All of them are in rows with Churn “No”.
- Remaining slides explain how the Data Set is exploited to arrive at solution.
- A note: Python and its libraries were used for Proof of Concept (POC).

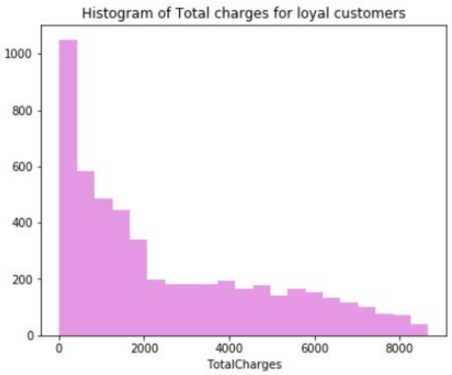
Data Stories (1) : Tenure in Months

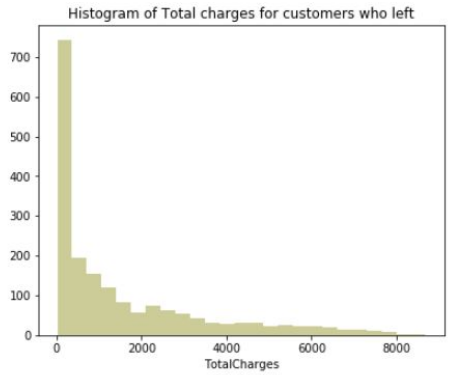

Histogram of tenure values for loyal customers

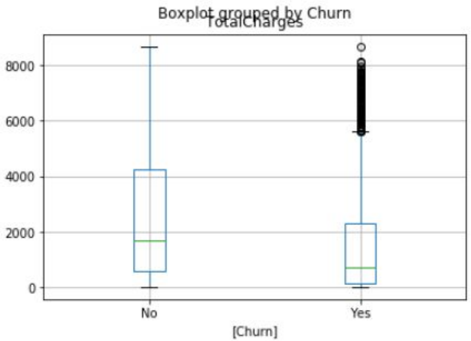

Histogram of tenure values for customers who left

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

Data Stories (2) : Total Charges in Dollars

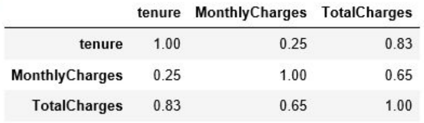




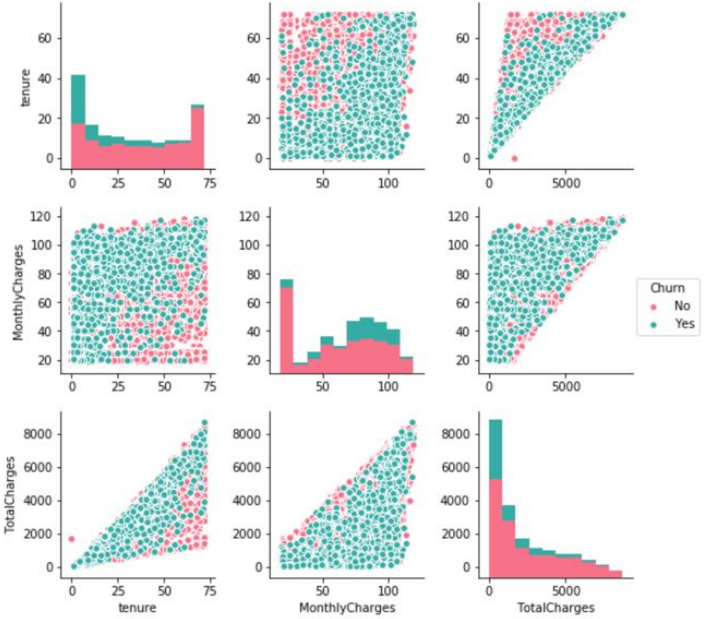


- Data Imputation: Missing values of Total Charges are filled with Median of Churn = ‘No’ group.
- Median (Churn = Yes) = $704
- Median (Churn = No) = $1684
- Median makes more sense due to skewed distribution
- Observe that loyal customers pay more total charges! Not counter-intuitive. Why? High tenure

Data Stories(3) : Correlation Plot

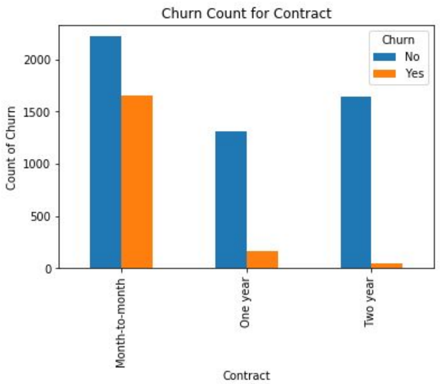


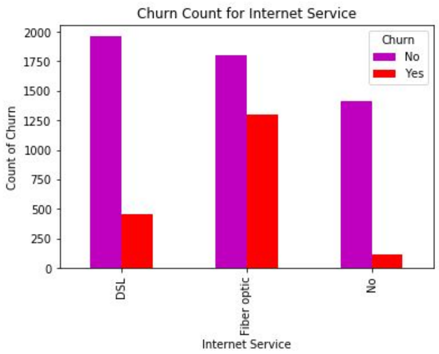
	tenure	MonthlyCharges	TotalCharges
tenure	1.00	0.25	0.83
MonthlyCharges	0.25	1.00	0.65
TotalCharges	0.83	0.65	1.00

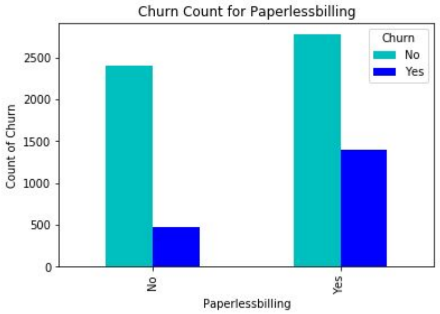


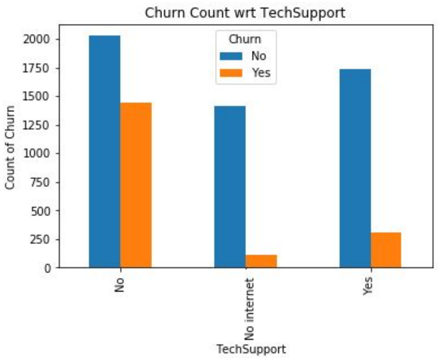
- Total Charges increases linearly with Tenure.
- The rate of increase is different for different customers: Fan effect.
- The correlation value 83% could result in skewness of some predictive models.
- Monthly charges and Total charges are correlated to 65% extent. This is also undesirable

Data Stories (4) Categorical Data









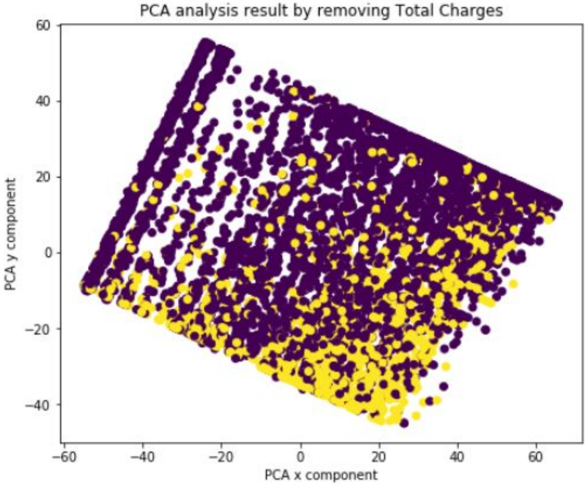
- There are notable differences in the counts and proportions of customers those switched.
- The Bar charts are self explanatory.
- Some more (not shown) below:
- Two genders' Churn behaviors don't differ much
- No Tech support, lack of online security => Less Loyal
- Customers with Partners and Dependents => More loyal
- Streaming => Slightly more loyal
- Automatic payment => More loyal
- The list is not exhaustive

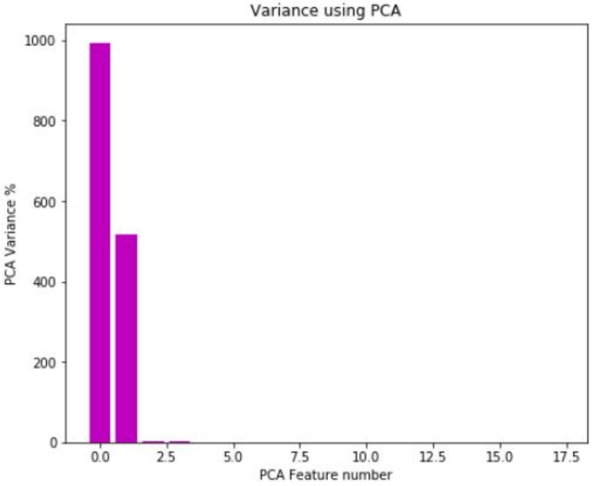
Predictive Modeling

- Data Prep/Feature Selection :
  - Drop numerical variable Total Charges (discussed already)
  - Drop the column Phone Service, because it is a subset of Multiple Lines
  - Convert to dummies, see the correlations and drop some before logistic regression
  - Do standardization before logistic regression
- Metrics used:
  - Mainly Area Under ROC Curve (AUC)
  - Accuracy on the test set.
  - “Recall” on the Churn class was given high priority for trade off

Principal Component Analysis (PCA)

- A quick PCA was done to get an idea of the nature of features
- It is found that 99% of the variance is explained by tenure and monthly charges

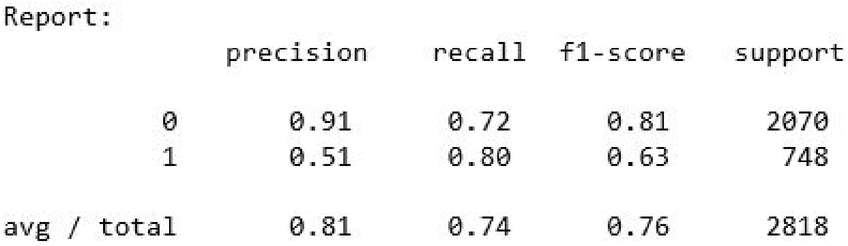




Logistic Regression

- Trained using 60% of the records,
- Membership from both classes in same proportion as original set
- Class weight Balancing done
- No hyperparameters
- The model gave 75% accuracy on both training and test set.
- The percentage of the times it predicted churn customers correctly was 80%
- The f1 score - 0.76, AUC was 0.84, indicating good fit.

Confusion Matrix	Pred 0	Pred 1
True 0	1497	573
True 1	146	602

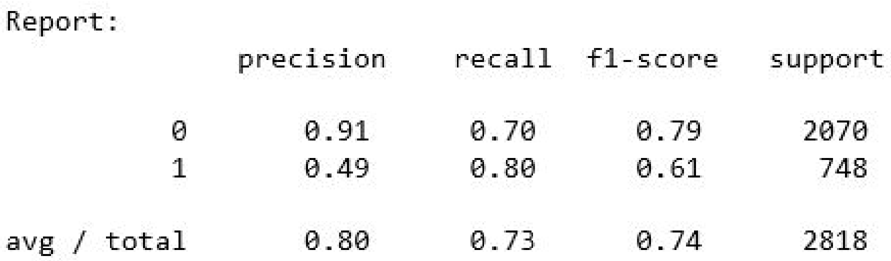


```
Report:
             precision    recall  f1-score   support

          0       0.91      0.72      0.81      2070
          1       0.51      0.80      0.63       748

avg / total       0.81      0.74      0.76      2818
```

Random Forest Classifier

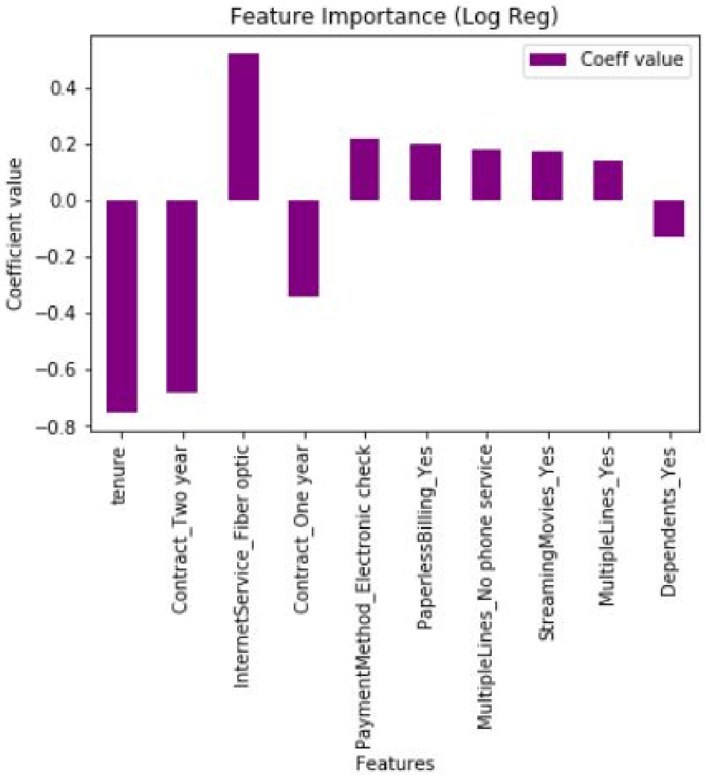
- First three points same as prev slide
- Some arguments were different from default
- Number of estimators: 25
- ‘Entropy’ index
- 4 features were used at each tree.
- Gave 73% accuracy on test set and 74% on training set
- The percentage of the times it predicted churn customers correctly was 80%
- The f1 score - 0.74, AUC was 0.84, indicating good fit.

Confusion Matrix	Pred 0	Pred 1
True 0	1456	614
True 1	149	599



```
Report:
             precision    recall  f1-score   support

          0       0.91      0.70      0.79      2070
          1       0.49      0.80      0.61       748

avg / total       0.80      0.73      0.74      2818
```

Feature Importance



- Top 10 Features with their influence
- Features with -ve coefficients are favorable to business
- How to interpret:
  - Electronic check, paperless billing are not friendly because they make churn easier
  - Streaming movies or the like for ex is not favorable if good service is not given
  - Should attract more customers with family.

Solutions Proposed

- Use the predictive models to identify the customers with higher inclination to switch to competition
- Take the following actions immediately:
    - Reduce charges. Although it might hurt short term revenue, good for long term
    - Improve on the Technical support on all services like streaming, phone connection and internet.
    - Encourage DSL, or improve Fiber optic internet.
    - Motivate sales guys to convince the customers to get into >= 1 year contract
    - Be up-to-date with current technology.
    - Collect customer feedback and act on it immediately to prevent new customer churn
    - Although billing/payment methods play a role, do not encourage customers to take paper billing or payments. It is not environmental friendly and will spoil the reputation.
- Next: It will be helpful to understand why churn started 5.5 years ago. Give more historical data to the data scientist for analysis.

Scope for Further Work

- Some K-Means clustering was done and quick conclusions were made. Split the records as per the K-means clusters, customize the incentives w.r.t. some other important factor might be possible.
- Tune the hyperparameters in predictive models. In the interest of time, only a basic tuning was done.
- Collect more data through surveys, analyze them using NLP techniques.
- Collect more historical data on customer churn. Especially it might be useful to see what happened 5+ years ago as seen from tenure plot.

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. My Code