Churn Analytics

Clustering and Classification

By Karan Singh

Outline

- Motivation
- Data
- Data Inspection and Treatment
- Exploratory Data Analysis
- Algorithms
- Feature Selection
- Modeling Results
- Model Interpretation & Use
- Churn Analytics Dashboard
- Next Steps

Motivation/Objective

- Customer Churn refers to the rate of customer attrition in a company or the speed at which customer leaves your company or service.
- Churn modeling is an important data science use case across many industries (especially subscription based businesses)
- Model churn and apply to a business situation in the telecom industry
- Learn how to perform clustering and classification
- Identify churn rates by important drivers of Churn

Business/Use case

Reducing customer churn by identifying potential churn candidates beforehand, and take proactive actions to make them stay.

Why is Churn a problem in telecom?

- One of the biggest pains in the telecommunications industry
- Average service provider in a mature market typically spends 15-20% of service revenues on acquisition and retention activities (Tefficient)
- Few new customers in mature markets, service providers must acquire them from their rivals.
- With service providers chasing the same group of out-of-contract customers, the Subscriber Acquisition Cost (SAC) of recruiting new customers is rising.
- Canada's BCE and Telus revealed that it cost almost 50 times less for them to keep an existing customer than to acquire a new one, with retention costs of C$11.04 and C$11.74 respectively, while average SAC in Canada weighed in at a whopping C$521 (2017)

Data

- Obtained from Kaggle: Telco Customer Churn
- 7043 entries, 21 columns (18 categorical , 3 numeric)
- Each row represents a customer, each column contains customer's attributes described on the column Metadata.
    - Customers who left within the last month – the column is called Churn
    - Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
    - Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
    - Demographic info about customers – gender, age range, and if they have partners and dependents

Data Inspection & Treatme

```
corr['Churn_Yes'].sort_values(ascending=False)
```

```
Churn_Yes                         1.000000
Contract_Month-to-month           0.405103
OnlineSecurity_No                 0.342637
TechSupport_No                    0.337281
InternetService_Fiber optic       0.308020
PaymentMethod_Electronic check    0.301919
OnlineBackup_No                   0.268005
DeviceProtection_No               0.252481
MonthlyCharges                    0.193356
PaperlessBilling_Yes              0.191825
StreamingMovies_No                0.130845
StreamingTV_No                    0.128916
StreamingTV_Yes                   0.063228
StreamingMovies_Yes               0.061382
MultipleLines_Yes                 0.040102
gender_Male                      -0.008612
MultipleLines_No                 -0.032569
DeviceProtection_Yes             -0.066160
OnlineBackup_Yes                 -0.082255
InternetService_DSL              -0.124214
Partner_Yes                      -0.150448
Dependents_Yes                   -0.164221
TechSupport_Yes                  -0.164674
OnlineSecurity_Yes               -0.171226
TotalCharges                     -0.198324
tenure                           -0.352229
```

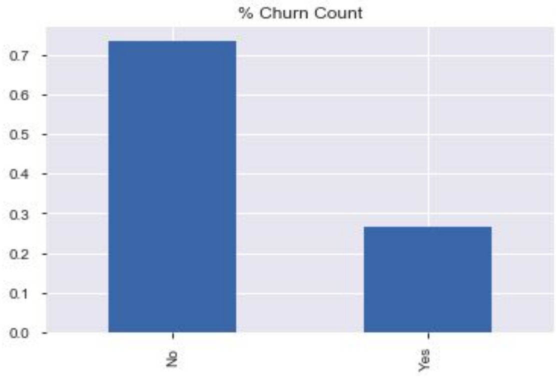
- Missing Values
- Outliers
- Categorical variables were encoded
- Pandas Profiling Report
- Pairplots
- Correlation Matrix

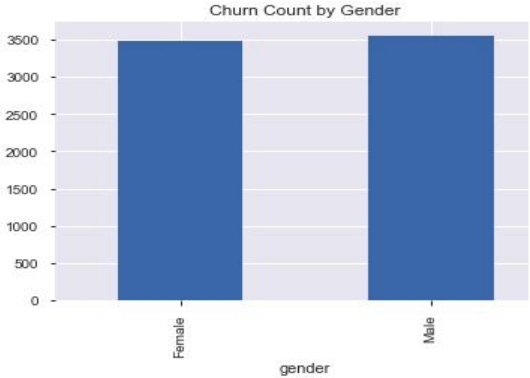
```
data.head()
```

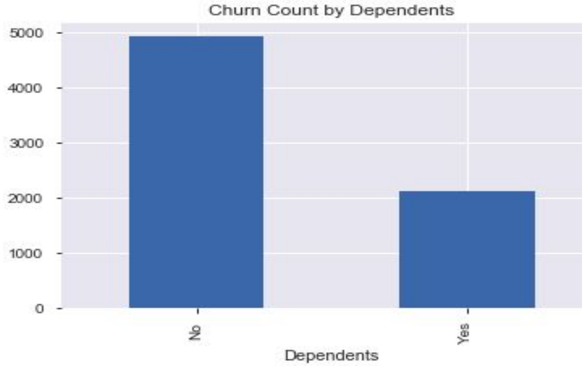
	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic

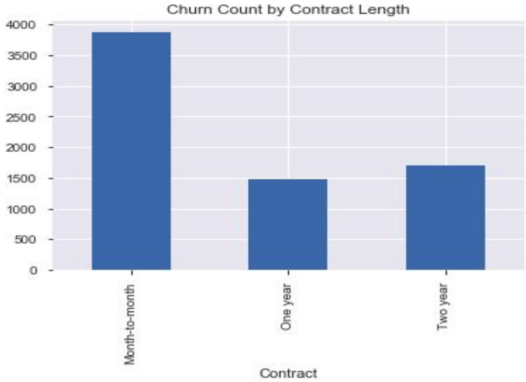
5 rows × 21 columns

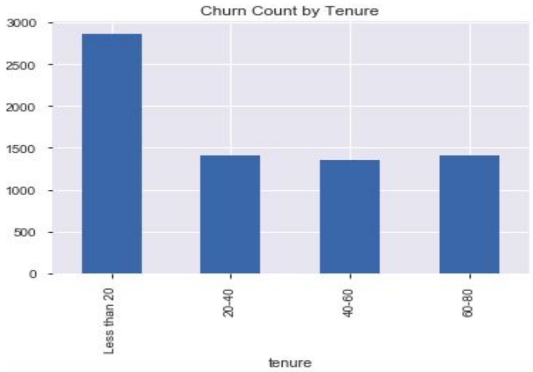
Exploratory Data Analysis

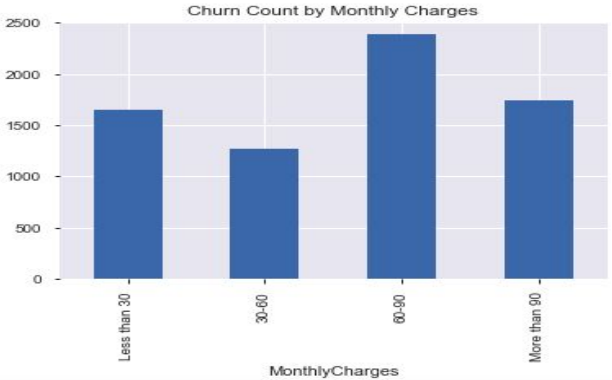












Algorithms

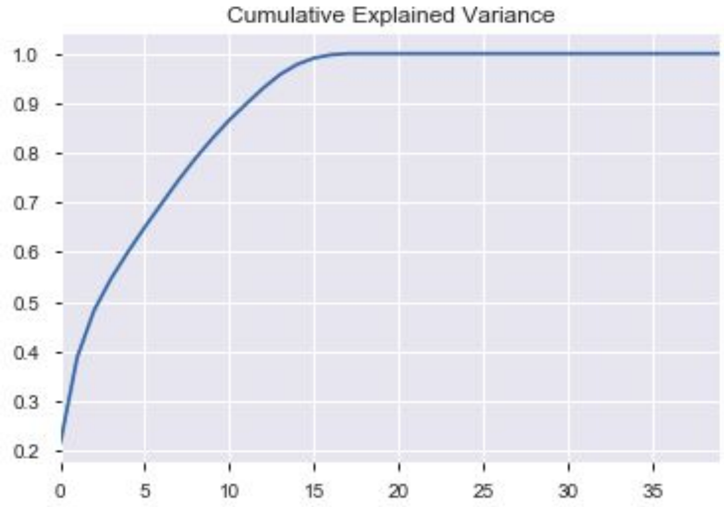
- **KMeans** (Clustering Analysis) & **Logistic Regression** (Classification)
- Predicting churn is a binary classification problem
- Based on evaluation metrics
    - Accuracy score does not work in an imbalanced dataset
    - Compared f1 and roc_auc scores for 4 algorithms: XGBoost (Baseline), Random Forest, Decision Tree and Logistic Regression
    - Highest roc_auc and f1 scores for Logistic Regression
- Along with being a robust model, Logistic Regression provides interpretable outcomes too

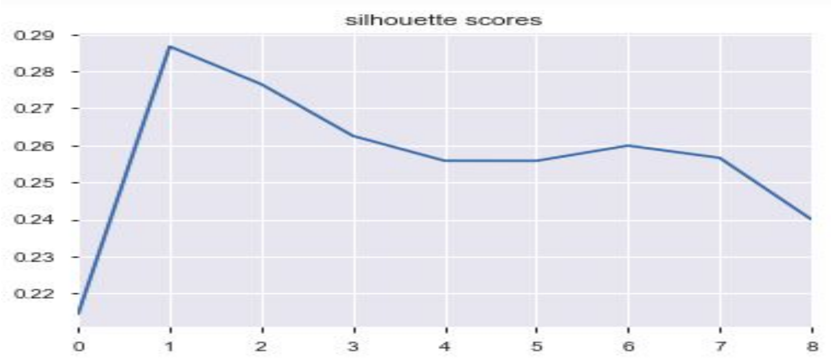
Clustering

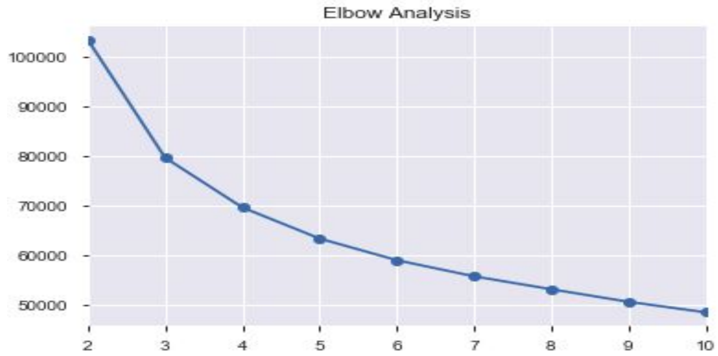
- Clustering is the task of gathering samples into groups of similar samples according to some predefined similarity or dissimilarity measure (such as the Euclidean distance)
- Some common applications of clustering algorithms include:
  - Primarily used for exploratory data analysis and business applications like customer segmentation, product segmentation, market segmentation.
  - Compression, in a data reduction sense
  - Can be used as a preprocessing step for recommender systems
  - Grouping related web news (e.g. Google News) and web search results
  - Grouping related stock quotes for investment portfolio management
  - Building customer profiles for market analysis
- KMeans (Based on Euclidean distance)

Clustering

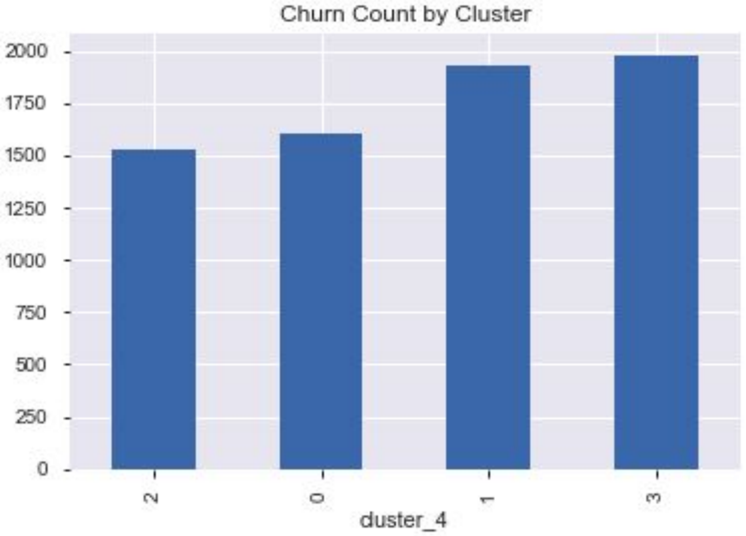
- Treated the business problem as an unsupervised business problem (removed Churn column)
- Using PCA: reduced dimensions from 40 to 8 explaining 75% of the cumulative variance
- Using KMeans to fit on the reduced dimensions and to obtain clusters

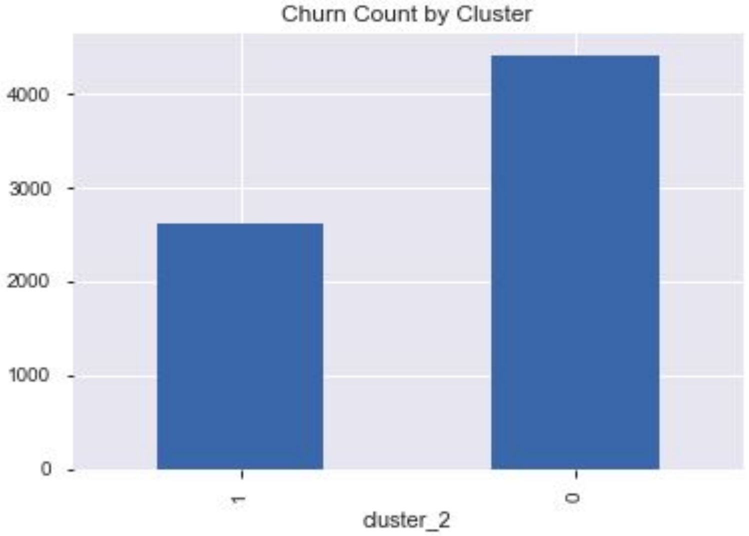






Engineered Features : Clusters





Classification : Logistic Regression

- Data split into 3 subsets : Train, Validation, Test set
- Model trained on Training set first and evaluated on Validation set
- Final Model trained on Training & Val and evaluated using Test set

```
pipe_final = Pipeline([
    ('scaling', StandardScaler()),
    ('oversampler', RandomOverSampler(random_state=42)),
    ('logreg', LogisticRegression(random_state=42))
    ])
params = {'logreg__C': [0.001,0.01,0.1,1,10,100], 'logreg__penalty':['l1', 'l2'] }
```

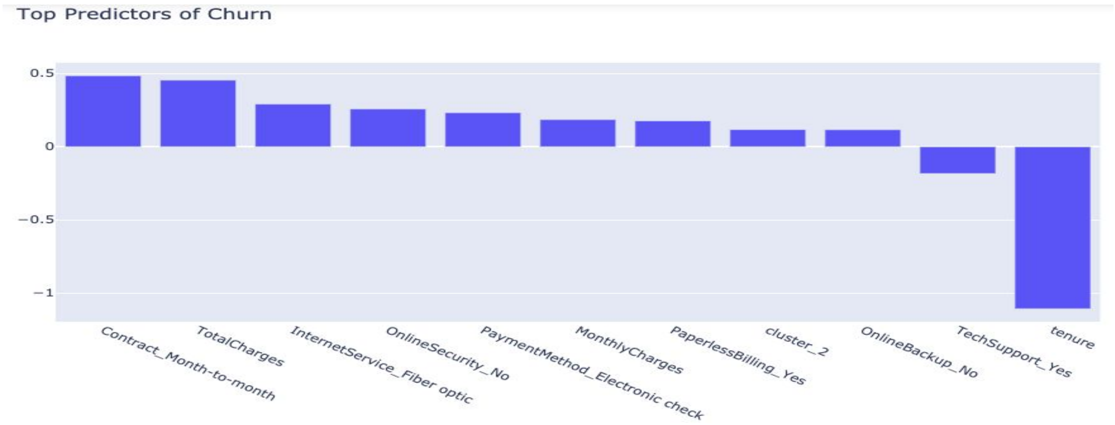
```
grid_final = GridSearchCV(estimator=pipe_final,
                          param_grid=params,
                          cv=5,
                          refit=True,
                         verbose= -1,
                         n_jobs=-1)
```

```
print(classification_report(y_test,y_pred_final))
```

	precision	recall	f1-score	support
0	0.91	0.71	0.79	1035
1	0.50	0.80	0.61	374
accuracy			0.73	1409
macro avg	0.70	0.76	0.70	1409
weighted avg	0.80	0.73	0.75	1409

Feature Selection

- Started with 42 columns (after dummy encoding)
- Features dropped progressively based on:
    - Correlation Matrix: Features with correlation less than 0.1 with the target variable were dropped
    - Use of wrapper algorithms
        - Logistic Regression, Random Forest, Decision Trees
        - Dropped features with minimal feature importances
    - Final set of features reduced to 10+1(cluster)



Modeling Results

- Top Features
  - Contract Type (Month to Month)
  - Total/Monthly Charges
  - Tenure
  - Internet Service (Fiber Optic)
  - Payment Method (Electronic Cheque)
- Recall score is high but precision is low
  - Determined with high certainty (80%) the actual churners
  - Misclassified many (50%) out of all the predicted churners (many False Positives)
- Trade off between number of features and metrics scores (more features generally imply a higher accuracy score)
- Clustering features (n=2 and n=4) turned out to be of low feature importance

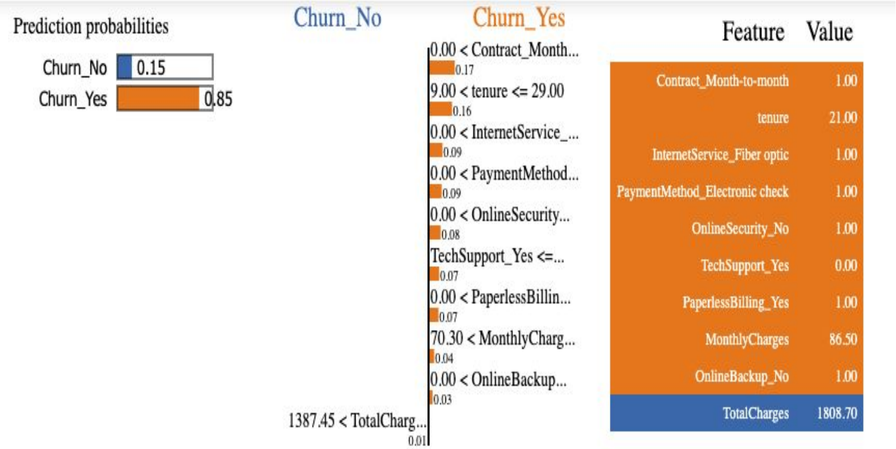
Model Interpretation & Use : Using Lime

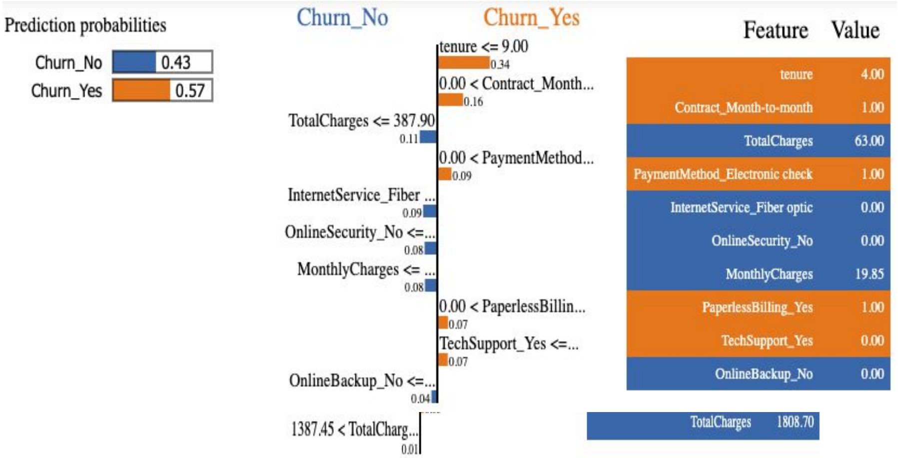
Test Case: customerID = '2057-ZBLPD'

y_test[‘Churn’] = 1

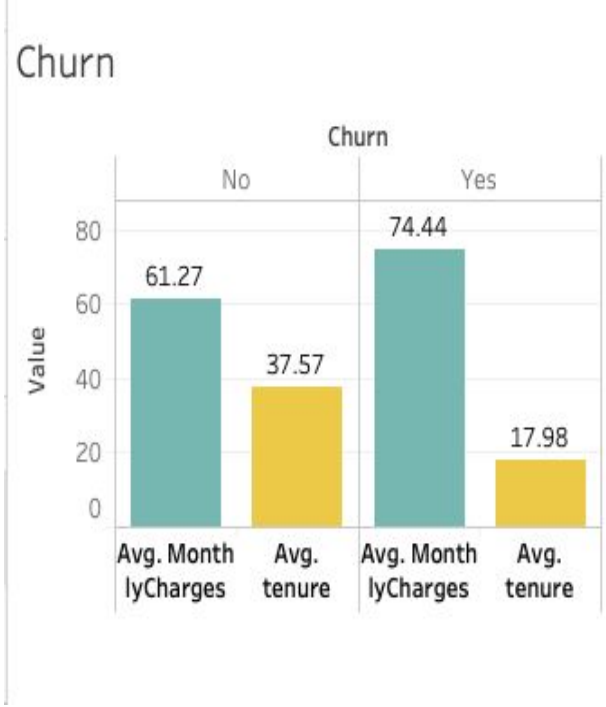
Test Case: customerID = ‘4396-KLSEH'

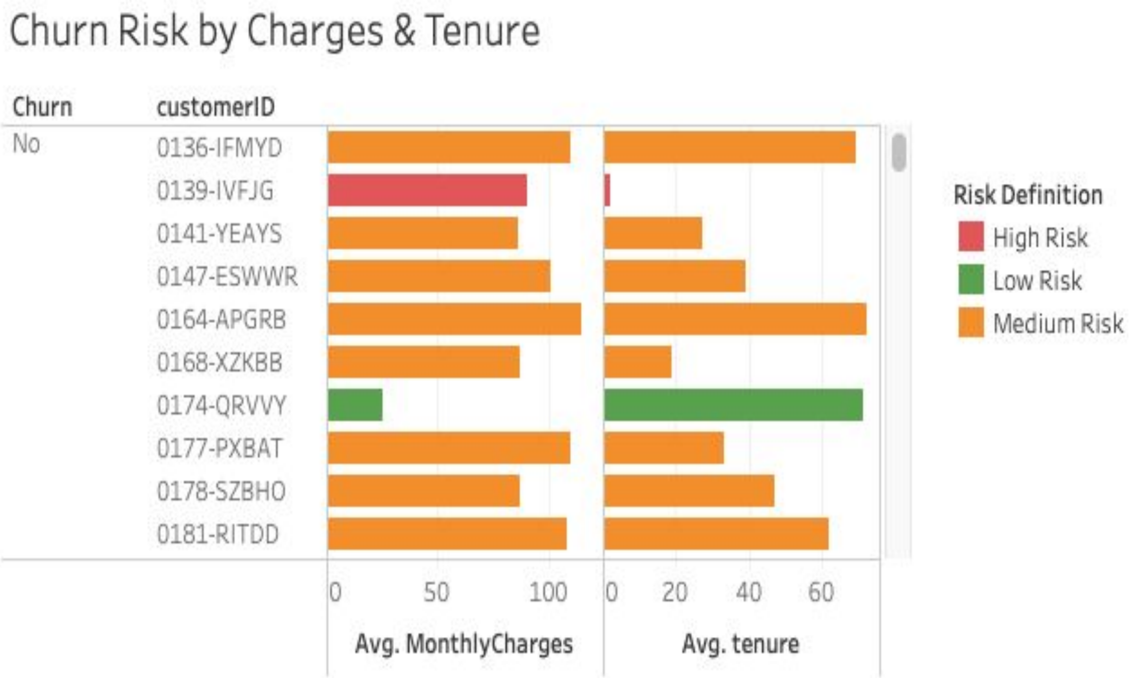
y_test['4396-KLSEH'] = 0



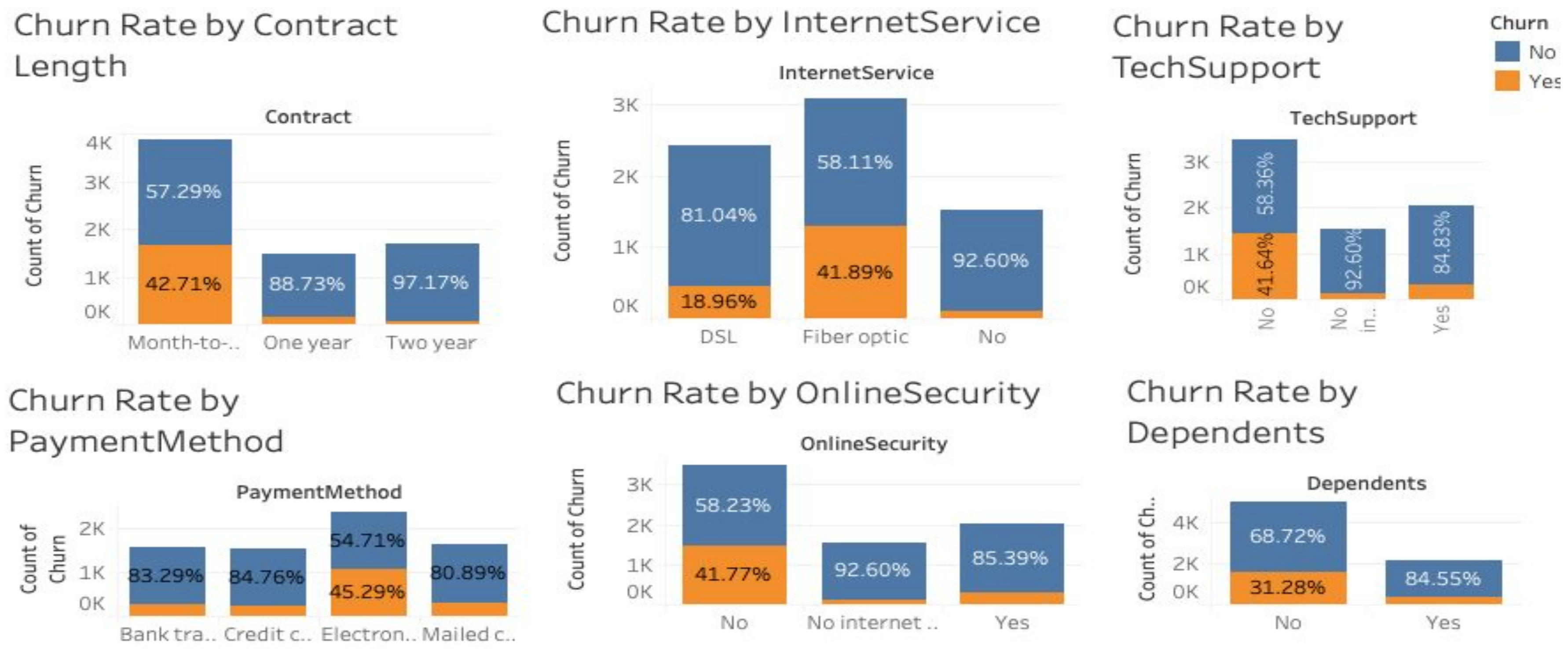


Churn Dashboard (In Progress)





Churn Dashboard



Challenges/Next Steps

- Challenges with respect to selecting the right dataset
- Further improve the model accuracy (especially precision scores)
- Use statistical methods for feature selection
- More feature engineering
- Try advanced algorithms
- Improve Dashboard

Thank You