Churn Analysis with Telco Customer data

Julia (Xueqing) Wang

Introduction

Customer Churn is when existing customer stop using the service/product provided by the company.

As it costs more to acquire new customers than it does to retain existing customers, it is important to understand the reason why customers churn and to implement tailored promotional plans to retain those high-risk customers.

In this project, I used the sample customer dataset from a telecommunication company, Telco, and applied python packages and machine learning models to analyze the churn problem.

Data Source

IBM sample dataset with 7043 observations and 22 variables that contain information about:

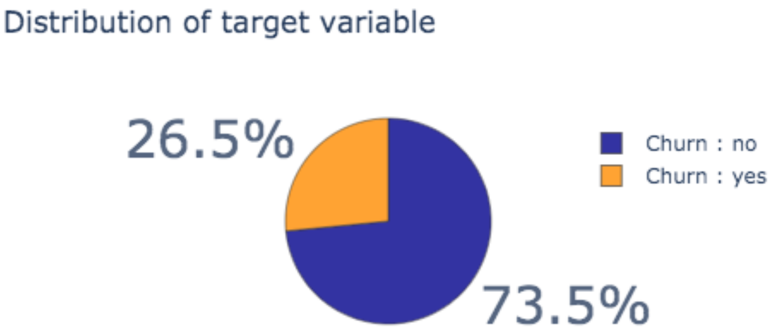
- Demographic info
- Services that each customer signed up for
- Account Information
- Churn: customers who left within the last month

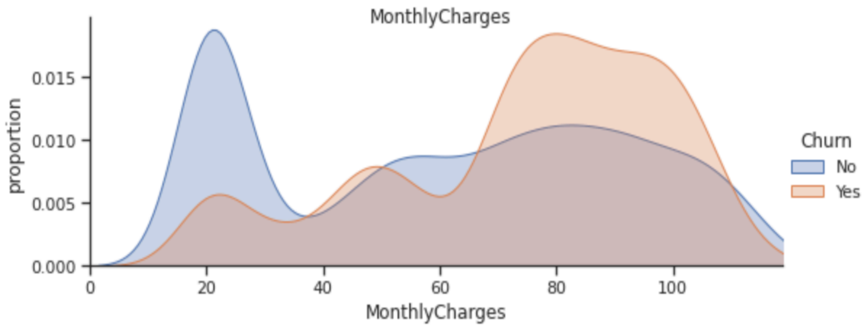
Data Preparations

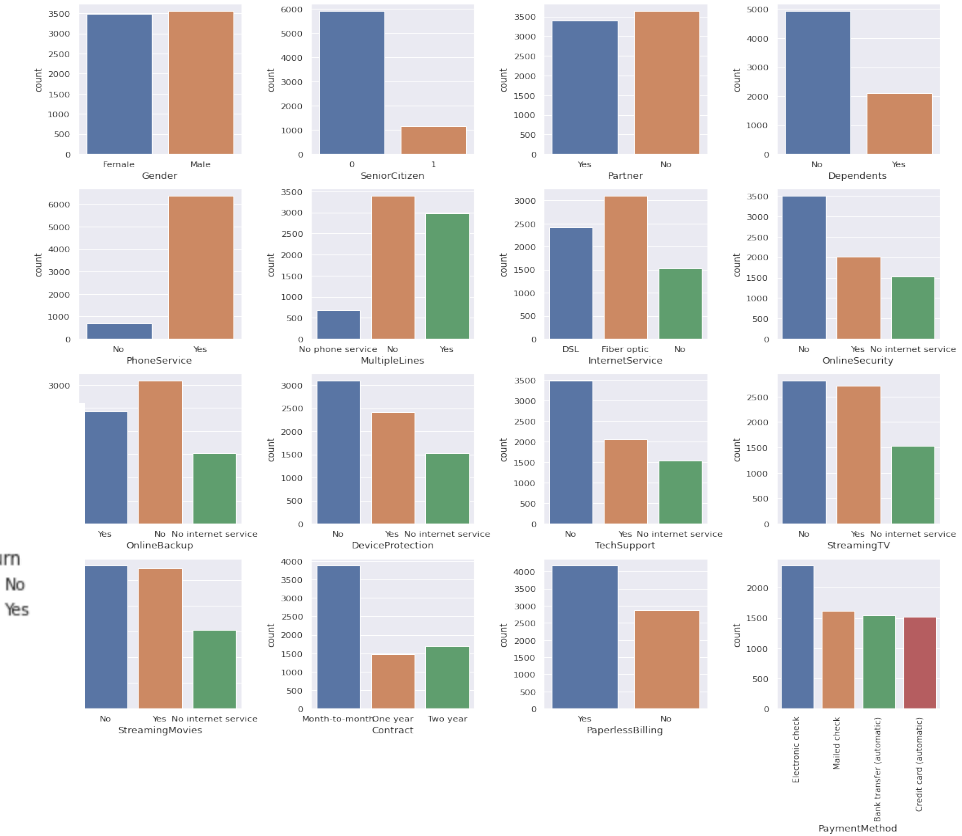
- Inspect the dataset
- Convert column data types and format the columns
- Identify the missing data and set them equal to the values in 'Monthly Charges'

*All the missing values are in 'total charges' column and all have 0 in tenure column, which means they are new customers and values of total charges equal to the values of monthly charges.

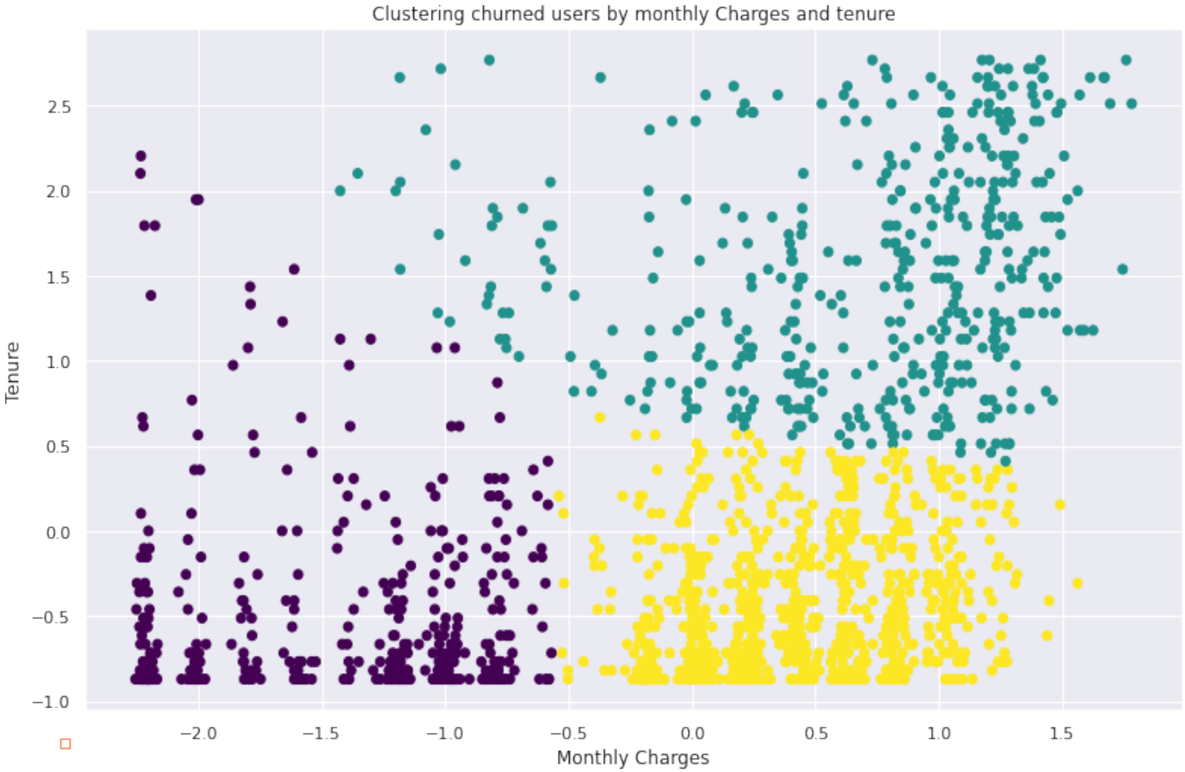
Exploratory Data Analysis







EDA K-means Clustering



- **P type (purple) Customer:**

  Low monthly charges(concentrated) and mostly short tenure, temporary customer looking for minimum services

- **Y type (yellow) Customer:**

  High monthly charges and short tenure **(tight distribution)**, price sensitive customers who churned quickly for cheaper services.

- **G type (green) Customer:**

  Long tenure(concentrated) and mostly high monthly charges, customers who stayed for a while because they either thought the service worth the price or due to lack of better alternatives.

Exploratory Data Analysis

Key points

- Categorical, numeric columns and target variable distribution was explored using seaborn,plotly
- K-means clustering was performed on tenure & monthly charges and three clusters, **one tight and semi-loose clusters** were created basing on the elbow method
- Promotion strategies should be tailored to each of churn customer type, especially to type G customers as they didn't immediately churn despite the high price

Data Modeling

Preprocessing

Get dummies manually for the categorical columns to make the data stay as clean as possible(also can be done using one-hot coding, and standardized the numeric columns

Building Models

Logistic Regression, Random Forest, SVM

Hyper Parameter Tuning

Tuned model hyper parameters using grid search

Selecting the model

Logistic Regression was selected for its best score on tuning process and for giving the highest AUC score

Model Interpretation

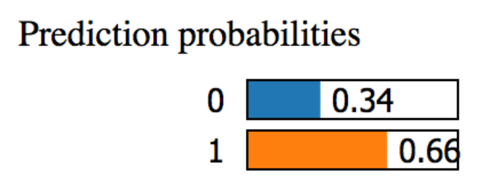
Why Model Interpretation matters?

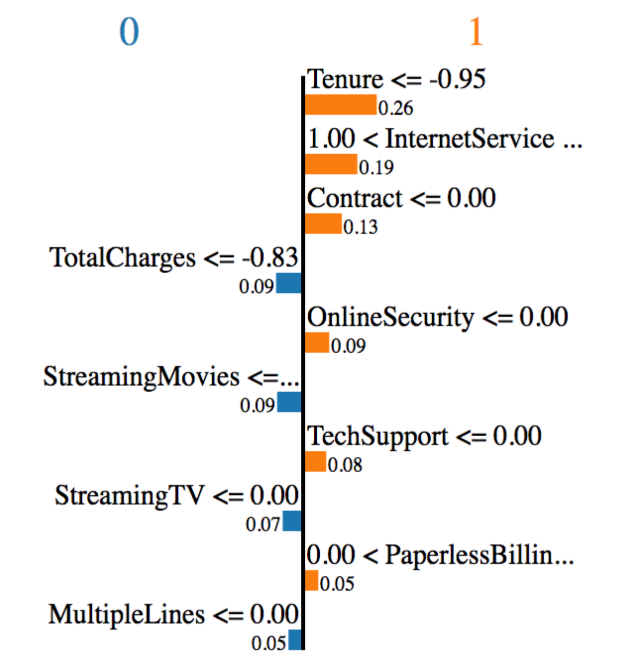
- If we can generate insights and explain the churns , we can reduce it
- We can transform the insights into actionable plans for the marketing team

Interpretation Methodology – LIME

- The Local Interpretable Model – Agonistic Explanation was selected to interpreted the model selected.
- It can explain individual predictions, enables analyst to select observation of interests

Use Case – Customer A

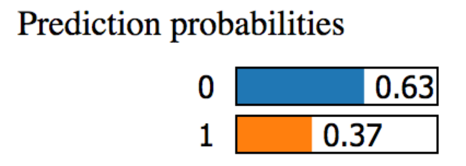




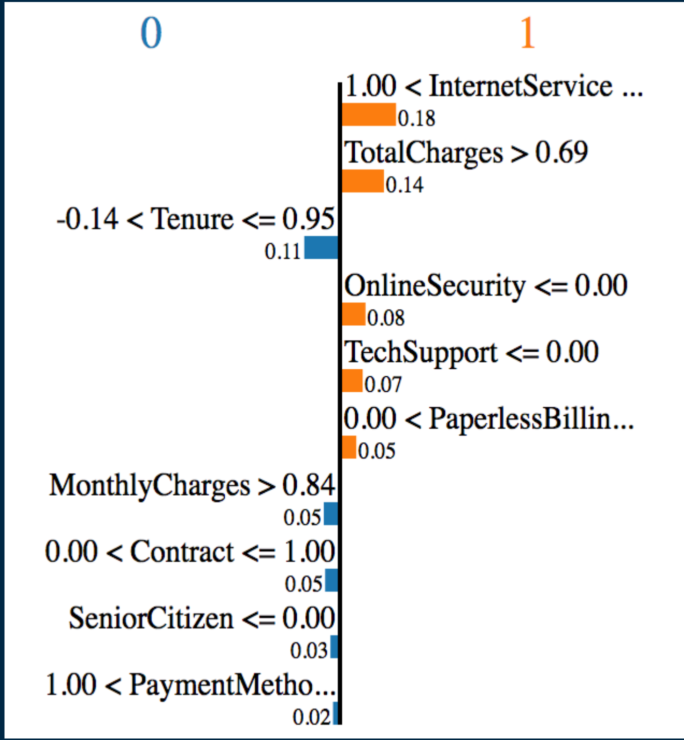
Feature	Value
Tenure	-1.28
InternetService	2.00
Contract	0.00
TotalCharges	-0.97
OnlineSecurity	0.00
StreamingMovies	0.00
TechSupport	0.00
StreamingTV	0.00
PaperlessBilling	1.00
MultipleLines	0.00

Tenure: -1.28 – one month
Internet service : 2 – fiber optic
Contract : 0 – month to month
Total Charges – 70.7
Paperless Billing - Yes

Use Case – Customer B



Internet Service: 2 – fiber Optic
Total Charges: 0.71 - 3886.45
Tenure : 0.15 - 36 (3 years)
Paperless Billing - Yes
Monthly Charges : 1.33 - 104.80
Contract : 1 – one year



Feature	Value
InternetService	2.00
TotalCharges	0.71
Tenure	0.15
OnlineSecurity	0.00
TechSupport	0.00
PaperlessBilling	1.00
MonthlyCharges	1.33
Contract	1.00
SeniorCitizen	0.00
PaymentMethod	2.00

Insights from analyzing Customer Samples

Customer A is a new customer and is 66 % likely to churn. This result is attributed to the short tenure(only one month), internet service type, contract and total charges.

Customer B has been a customer for 3 years and is 37 % likely to churn. The fact that this customer is paying 3886.45 of total charges and internet service has increased the likelihood of churn, the model decided customer B is more like to stay due to the relatively long-term contract and tenure.

Summary

Insights and Recommendation

- Group of customers with high monthly charges and short tenure most likely to churn
- Month- to- Month contract customers are churning fast
- G type customer should be the prioritized target customers since they are more likely to be retained
- Offer customized promotions and apply A/B test on promotion strategies

What to do next?

- Further investigated and keep improving the model
- Apply hypothesis tests to find the non-effective attributes
- Analyzed the effects of different internet service type