**Thought Process**

My chosen dataset is a customer data from a telecom company. The main objective is to predict which customers will stop being customers (Churn) and those that will remain. From early data exploration. There were 9 missing entries in the “Total Charges” columns. This either indicates that the entries are truly missing or they are actually 0 values. I went for the option of imputing the missing values with 0 because there might be a chance that those customers actually started their first month of service. Hence, they are not required to pay yet.

Data Preparation and some Basic EDA

The next step is interpreting the data and checking for duplications. There are 6499 rows and 21 Features. “CustomerID” was dropped because they are all unique ID values which doesn’t contribute to analysis processes. I ran a for loop to look for the frequency of each categorical features instead of running the common procedure of plotting histogram. The outcome values “Churn” appeared to be imbalanced. Afterwards, describing numerical features then onto feature engineering and selection.

Feature Engineering and Selection

Correlations test was run and there was a high correlation between Tenure and Total Charges. This is expected since as Tenure increase, Total Charges is also expected to increase accordingly. A simple calculation task was conducted to see if Total Charges Column is the product of Tenure and Monthly Charges and turned out it was actually the case. However, the approximated calculation values and the actual value were similar but not the same, which might indicate that there is a scenario where the customers either get a discount or have to pay a fine for late payment. “Total Charges” column is dropped because it is high-correlated with Tenure. Label Encoding was conducted next and then a chi-square test. “Phone Service” and “Gender” were dropped because they exceeded the 5% significance level threshold. Then, I oversampled the dataset to create a balanced dataset for modelling.

Modelling

Dataset then was split into train set and test set. Logistic Regression and Tree-based models came to my mind since they could predict binary variables. Naïve-bayes would work pretty well too If I discretize numerical features. Since the dataset was mixed, distance-based models were excluded. Finally, I chose random forest over boosting and other techniques since Random-forest is easier for parameter tuning over boosting technique and it is slightly better at working with large dataset than Logistic Regression and Naïve-Bayes. Boosting Techniques would work on Imbalanced dataset pretty well. However, I wanted to try out Random Forest.

Model was run, returned results seemed good. However, tree-based models are prone to overfitting, so I decided to tune and cross-validate to find the best parameters. Predictions and confusion matrices were conducted on all three Training, Testing and Un-sampled dataset. Afterwards, a random forest’s feature importance table was made for later Tableau visualisation.

Predicting Real World Dataset

I repeated all the steps of dropping columns, encoding and prediction using earlier random forest model. Since the Real-World dataset doesn’t have the outcome variables. We could only do prediction and compare it training dataset.

Clustering on customers who churned for re-marketing purposes

Predictions are great but business value often come from unsupervised learning task. I intended to cluster customers who has left the services to see the differences between each group. The process of encoding, interpreting and dropping irrelevant features were repeated. Since the dataset was mixed, distance-based clustering techniques such as k-Means were no good. Hierarchical clustering with Gower distance and K-Prototypes came to my mind. I went for the latter since Hierarchical clustering usually take much longer to compute. I ran a for loop on a range of values of number of clusters to find the optimal number of clusters with Silhouette Score. Optimal number of clusters then was found to be 3. Then a tree graph was plotted to interpret the properties of each clusters.

**Suggestions and Key Insights**

Contract Type, Monthly Charges and Tenure are important features for explaining Churn-rate.

Month-to-month customers have a usually high churn rate at the 70$ - 110$ Monthly Charges Bracket. Fluctuating about 49.82% to 53.94%. Surprisingly, these customers also make the most money for the company. Good news is that they are gradually less likely to leave as their tenure increases.

* Provide Monthly Charges discounts as their tenure increases. Therefore, they would hopefully give in to the promotions and stay longer and thus generating more profit.
* There also might be a chance that Telecom competitors provide a better high-end service at a high Monthly Charges. Thus, customers might switch.

Churn-rate of Customers on one-year or two-year contract gradually increase over-time however not at a significant rate and not as high as Customers on Month-to-month contract.

* Try to make Customers on Month-to-month contract to switch to longer plans. Promotions like discount could be ideals for longer plans.

Monthly Charges of Customers on one-year contract drives up churn-rate gradually increase over-time from 2.76% to 22.92%.

* Another proof of competition at high-end telecom service.

Green cluster: Monthly Charges < 60.225 and Tenure < 27.5 (Account for ~24% of the Total Churners)

* Only 24%, which means the company does reasonably well at non-premium bracket.

Orange cluster: Monthly Charges > 60.225 and Tenure < 23.5 (Account for ~49% of the Total Churners)

* Another proof of competition at high-end telecom service. However, Tenure less than 23.5 indicates that customers will be less likely to leave as Tenure increases. This clusters accounts for a total of ~49% of the Total Churners as well, which is a strong indication that this telecom company needs to improve their service for their premium customers. This bracket produces the most profit as well, high churn-rate at this bracket means losing profit. It is suggested to aggressively target this group for marketing and re-marketing purposes, trying to get customers on this bracket to stay longer is ideal.

Violet cluster: Monthly Charges > 80.225 and Tenure > 27.5 (Account for ~19% of the Total Churners)

* Another proof of competition at high-end telecom service. However, Tenure is greater than 27.5 and it accounts for 19% of the total churners, which is OK. Since as tenure increases, people are less likely to leave.

**What can be improved?**

Time-series is needed for the dataset. Therefore, we could know which month customers are likely to leave, projected profit, … Seasonality factors plays an important role in predicting churn-rate such as in Job jumping period (Work phone might be switched) …

Earlier we founded that Total Charges Column is the product of Tenure and Monthly Charges. This means that customers might have taken discount promotions or have paid fines due to various reasons. These variables need to be included in the dataset. Number of taken promotions might have an impact on whether the customers are going to continue to stay with the company. Number of times a customer has to pay a fine due to late payment might have an impact on whether the customers are going to not continue with the company.

List of competitors pricing at different bracket, and the kind of features they provide. This would help with pricing strategies and optimization.

More attributes of customers can also be collected such as ages, did they recently switch telecom provider?, Location (Location can be visualised for Regional Profit as well), How much data have they used in the past? (per month, per year …). (Important since how much data they have used could predict whether they are going to end their contract or not) …

Customer sentiment analysis. Use survey to see whether the customers are happy with the service or not. Thus, we could predict churn rate based on those data.

Fraud detection. Are your customers use your telecom service to send spam messages and to make spam calls? Those who do this are likely to jump across different telecom providers (especially one that provides cheap value packages). Thus, detecting them is valuable to predict churn-rate.

**Keep calm and data science.**