Summary

Forecasting and comprehending customer retention is major concern in telecommunication industry. This is mainly due to intense competition and higher customer acquisition cost which makes acquiring new customer more expensive than cost of retaining existing customer base. Maximising each customer's lifetime value will lead to stable profitability in the business. The purpose of this investigation is to develop a binary classifier model that scrutinize the diverse factors affecting customer churn within the industry. A comprehensive analysis of historical customer data is carried out to identify the major reasons and likelihood of churning. Identify and focusing on these factors can help telecom industries to devise strategies and approaches to improve retention rate and increase customer loyalty. Further creating a positive brand image by providing personalized offerings, increases the value provided to customers which in turn offers a competitive advantage within the industry.

Moreover, CRISP – DM analytical methodology is utilized to train the machine learning models. Initially, data pre-processing and Exploratory data analysis is carried out to get the better understanding of the dataset. Then various classification machine learning models are trained by utilizing python and hyperparameter tuning and model evaluation is carried out to find the best models which are CatBoost., XGBoost, Random Forest and LGBM and 0.95 F1 score. Sentimental Analysis is also carried out to get the insight of the sentiments of the customer which can help to improvise the various services of the telecome industry.

Table of Contents

Summary *1*

1. Abstract *4*

2. Background *4*

3. Literature review and Previous work done *5*

4. Project Objectives *6*

5. Analytical Methodology: CRISP-DM *6*

5.1 Business Understanding **7**

5.2 Data Understanding **8**

5.3 Data Preparation **11**

5.4 Modelling **20**

5.5 Evaluation. **24**

6. Sentimental Analysis. *27*

6.1 Histogram Analysis: **28**

6.2 Identifying the factors causing negative sentiments. **29**

6.3 Comparison with Corelation results: **29**

6.4 Tools and Technologies used in the Project: **30**

7. Result and Conclusion *30*

8. Future Scope *32*

9. References *32*

Table of Figures

Figure 1: CRISP-DM methodology. 6
Figure 2: Churn label distribution 11
Figure 3: Data type of the attributes. 11
Figure 4: Total Charges column with null values 12
Figure 5: Data type conversion – Total Charges column. 12
Figure 6: Missing values in the Total Charges column 12
Figure 7: Replacing 0 with missing values – Total Charges column. 13
Figure 8: Checking duplicate data values. 13
Figure 9: Distribution of numerical variables. 13
Figure 10: Factors influence on churned and non churned customers. 14
Figure 11: Top Churn Reasons – Bar chart. 15
Figure 12: Customer churned based on contract type– Pie chart. 16
Figure 13: Customer churned based on contract type– stacked bar chart. 16
Figure 14: Customer churned based on contract type– Bar chart. 17
Figure 15: Churn reason comparison with CLTV – Bar chart. 17
Figure 16: Corelation between variables – Heat map 19
Figure 17: Data splitting – Code snippet 20
Figure 18: ROC Curve & positive Rate for all models 24
Figure 19: Best models F1 Scores 25
Figure 20: Best models in regards to hyperparameter 26
Figure 21: Confusion matrix for the catBoost 26
Figure 22: Top Churn reason 27
Figure 23: Sentiment analysis script 28
Figure 24: Sentiment analysis script 28
Figure 25: Distribution of Sentiment Score 28
Figure 26: Top Churn reason 29
Figure 27: Top Churn reason 30

1. Abstract

Churn is a major problem in many industries and companies are competing to achieve a low customer attrition rate. Telecommunication industry is one among the severely affected as its very easy for customers to switch from one to another due to the wide variety of options available in market. Identifying customers who are most likely to cancel their subscription can help business develop strategies to retain the customer base while increasing their revenue.
Customer churn can include customers switching from one operator to another or cancelling the subscription. It's important to identify customer at the risk of churning and improve their customer lifetime value (CLTV) to retain the subscription, this in turn increases the business revenue.

Here we use machine learning to create a model that predicts the customers that are at-risk of churn by analysing a set a features like user demographics, subscription pattern, pricing information etc. This will help the business to develop strategies to reduce the customer churn. Accurately classifying churn is a challenge due to the various number of factors that might influence churn and not always the business knows the reason for churn. We will be exploring different machine learning models and tune and compare them to get the best model. The dataset used is IBMs Telco Churn data set, which has customer details including customer demographics, subscription details, subscription status and if churned the reason. By using the Churn reason user sentiments can be analysed.

Further we make use of visualizations to identify the common reasons for churn and evaluate the areas that need to be addressed while creating retaining strategies and policies. The visualizations will also give insights to the current state of business and where it lacks. It can also serve as a guide to target customers by analysing behavioural pattern of the customers who did not churn

Our research emphasizes the significance of interpretability in machine learning models. We provide businesses with actionable insights by identifying the key factors that contribute to customer churn, allowing them to improve their services and prevent churn. We show how machine learning techniques can improve customer retention and reduce churn rates.

2. Background

IBM has conducted research to investigate telco customer churn and to address this challenge, IBM's research uses data analytics to analyse customer behaviour, including factors such as demographics, usage patterns, customer service interactions, and billing issues, among others. The objective is to identify patterns and factors that contribute to customer churn and to recommend strategies to improve customer retention. This may involve suggesting changes to pricing or packaging, improvements to customer service, or new marketing strategies to attract and retain customers. By reducing customer churn, telecommunications companies can improve customer satisfaction, loyalty, revenue, and profitability.

3. Literature review and Previous work done

Several studies have been conducted to explore the use of machine learning algorithms in Churn prediction for telecommunication industries, one of them is a customer churn prediction model developed using multi-layered NN to determine the possibility of churn and factors influencing. The model created achieved an accuracy of 80.03% and helped in finding behavioural patterns from an unstructured sheer volume of data. However, the model did not explore the sentiments of customers and possible churn reasons (Agrawal et al., 2018). A similar study by Momin, Tanuj and Raut used different classification algorithm to predict churn, this model could avoid the hustle of feature engineering by using deep learning and identified that Multi-layered ANN outperforms the traditional classification model due to its self-learning capability. Here the model failed to address the imbalance of the dataset (Momin, Tanuj Bohra and Raut, 2020). Authors Pamina, Sathyabhama, et al. tried to predict the influence of different customer traits using different classification algorithms and compare them. The study found XG Boost algorithm gave the best results and customer with fibre optic is found to be the highest influence on churn (Pamina et al., 2019). The model did not address the imbalanced data problem and did not focus on continuous learning, which can make the model better in long run. Effectiveness of Gradient Boosting algorithm and how over sampling affected the classifier performance was explored but the maximum accuracy that the classifier could achieve was 70%. Here the authors focused on a single algorithm, and this limits the potential of a classifier model to enhance its efficiency when combined as in case of ensemble techniques (Halibas et al., 2019). Further studies made use of clustering and association rule along with classification model to predict churn, this model could achieve the best results with back propagation and multilayer perceptron and used DB scan to cluster customers to churners and non-churners, Factors for churning were analysed using Apriori and FP growth algorithms. This paper paves the idea to use clustering techniques to label the less informative data (unlabelled data) and makes the most use of it. Even though the problem of imbalanced data is not addressed and there is limited feature selection (Mitkees, Badr and ElSeddawy, 2017).

Although these studies used machine learning algorithms and analysed various factors influencing churn. There is still a gap to analyse how combination of classifiers or ensemble method can enhance the model and how concepts like Active learning and Survival analysis can be implemented to create a better churn prediction model. The user sentiments and churn reason need to be analysed this helps the business identify its strengths and weakness and identify the nature of its customer base. This will aid the business to help make better informed decisions to increase customer satisfaction and retention rate.

4. Project Objectives

- To perform sentiment analysis on customer feedback to gain insights into customer perceptions and opinions about a product or service provided by the company.
- To identify the major contribution of factors influencing the customer churn.
- To create machine learning classification models to predict the possibility of customer churn.

5. Analytical Methodology: CRISP-DM

Figure 1: CRISP-DM methodology.

The methodology that needs to be selected to analyse the data mainly depends upon the research goal or business goal, type of data being considered and also the scope of the study. Amongst various methods, we believe CRISP-DM(Cross-Industry Standard Process for Data Mining) is the best to proceed with Telco Customer Churn dataset as phases of this methodology will help to understand objectives of the study, explore the data for understanding the factors leading to customer churn and also build predictive models focusing on customer retention and evaluate its performance.

The six phases of CRISP-DM mainly help to build a systematic and structured approach for data mining project.

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment

5.1 Business Understanding

Business understanding is the first phase of CRISP-DM methodology. This phase ensures effort for data analysis and modelling to be aligned with stated business goals. The phase is responsible for various tasks such as determining the business objectives, assessing the current situation of the company, determining the data mining goals and creating a project plan to for the business. For Telco company to determine the business goals, information regarding the products and services provided to the customers has been collected and percentage of churned and non-churned customers is gathered. Based on these data the business objectives have been set focusing on increasing the retention rate of the customers and their loyalty towards the company.

As the business objectives have been set as a next step this objective has been converted into data mining goals. One such goal is to perform the sentiment analysis on feedback provided by the churned customer based on products and services provided by the company. This data mining task will help to understand the customers perceptions towards company services and products. These factors will help to identify the area of improvement in services offered to customers and increase customer satisfaction and decrease customer churn rate. As customer behaviour pattern is identified, this technique will help the company to identify potential indicators of churn and develop tailored retention strategies, prioritizing efforts on high-value customers to maximize their long-term value to the company.

Further the list of different models that will be used to predict the churn possibility are Logistic Regression, Decision Tree, Random Forest, XGBoost, K-nearest Neighbour, Neural Network, AdaBoost, LGBM, CatBoost, Support Vector Machine. The reason behind choosing these models is because of their capacity to handle large number of input features as telco company has several factors to be considered as input feature to predict the churn rate and dealing with non-linear relationships where churn prediction data often involves capturing non-linear variables between input variables and target variables. These models are also capable of modelling this complex non-linear relationship by adopting a combination of weak learners such as decision tree and boosting techniques which allows these models to capture intricate patterns and interaction among features, improving predictive accuracy for customer churn.

The Business Understanding also deals with project plan where other phases such as cleaning of data, selection of right variables that can be used in supervised learning, model selection and model evaluation has been used in this project. Here the project also deals with quality of the data as the data provided by IBM is a fictional telecom data, there is a possibility of model being overfitted or underfitted due to data imbalance. To resolve the problem of imbalance in dataset, SMOTE resampling technique is used. This technique generates the minority classes equivalent to the number of majority classes by exploring the relationship in minority class. Hence eliminating the data imbalance in the available data.

5.2 Data Understanding

Data understanding phase involves acquiring initial understanding and gaining insights into characteristics, quality, and relationships among the variables of the dataset. The key features that have been involved in data understanding phase of CRISP-DM are as follows:

1. **Data Collection:**

   Once the business understanding phase is completed and the project objectives are well-defined, the next step is to proceed with data collection. Data collection involves identifying and gathering the relevant data sources that are necessary to address business questions and achieve project goals. This may involve accessing databases, retrieving data from different systems, or obtaining data from external sources.

   The dataset chosen is ''Telco Customer Churn Data" a fictional telecom company. The secondary dataset is available on Kaggle, an open-source data aggregation platform and the dataset is sponsored by IBM Community.

2. **Data Description:**

   In this section the detailed description of the Customer churn dataset including the data format and its content has been explained.The dataset provides insights into a hypothetical telecommunications company based in California, United States which offered home phone and Internet services to customers during the third quarter. The dataset includes information about customer Demographics, Customer Location details, Services provided by Telco to customers and customer churn status. The dataset is organized into rows and columns. Each row represents a unique customer and contains information on their various attributes.

   The dataset has 33 columns and 7038 rows, which represents the total number of customers included in the dataset.

Customer Demographics related columns and description:

Sl. No	Column	Column Description	Column Type
1.	CustomerID	A unique ID that identifies each customer.	String
2.	Count	A value used in reporting/dashboarding to sum up the number of customers in a filtered set.	Integer
3.	Gender	The customer's gender: Male, Female	String
4.	Senior Citizen	Indicates if the customer is 65 or older: Yes, No	Boolean
5.	Partner	Indicates if the customer has partner: Yes, No	Boolean
6.	Dependents	Indicates if the customer lives with any dependents: Yes, No. Dependents could be children, parents, grandparents, etc	Boolean

Customer Location detail related columns and description:

Sl. No	Column	Column Description	Column Type
1.	Country	The country of the customer's primary residence.	String
2.	State	The state of the customer's primary residence.	String
3.	City	The city of the customer's primary residence.	String
4.	Zip Code	The zip code of the customer's primary residence.	Numeric
5.	Lat Long	The combined latitude and longitude of the customer's primary residence	String
6.	Latitude	The latitude of the customer's primary residence	Float
7.	Longitude	The longitude of the customer's primary residence.	Float

Telco Service detail related columns and description:

Sl. No	Column	Column Description	Column Type
1.	Tenure Months	Indicates the total amount of months that the customer has been with the company by the end of the quarter specified above.	Integer
2.	Phone Service	Indicates if the customer subscribes to home phone service with the company: Yes, No	Boolean
3.	Multiple Lines	Indicates if the customer subscribes to multiple telephone lines with the company: Yes, No	Boolean
4.	Internet Service	Indicates if the customer subscribes to Internet service with the company: No, DSL, Fiber Optic, Cable	String
5.	Online Security	Indicates if the customer subscribes to an additional online security service provided by the company: Yes, No	Boolean
6.	Online Backup	Indicates if the customer subscribes to an additional online backup service provided by the company: Yes, No	Boolean

7.	Device Protection	Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company: Yes, No	Boolean
8.	Tech Support	Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times: Yes, No	Boolean
9.	Streaming TV	Indicates if the customer uses their Internet service to stream television programming from a third-party provider: Yes, No. The company does not charge an additional fee for this service.	Boolean
10.	Streaming Movies	Indicates if the customer uses their Internet service to stream movies from a third-party provider: Yes, No. The company does not charge an additional fee for this service.	Boolean
11.	Contract	Indicates the customer's current contract type: Month-to-Month, One Year, Two Year.	String
12.	Paperless Billing	Indicates if the customer has chosen paperless billing: Yes, No	Boolean
13.	Payment Method	Indicates how the customer pays their bill: Bank Withdrawal, Credit Card, Mailed Check.	String
14.	Monthly Charges	Indicates the customer's current total monthly charge for all their services from the company.	Numeric
15.	Total Charges	Indicates the customer's total charges, calculated to the end of the quarter specified above.	Float

Customer Churn Status related columns and description:

Sl. No	Column	Column Description	Column Type
1.	Churn Label	Yes = the customer left the company this quarter. No = the customer remained with the company. Directly related to Churn Value.	Boolean
2.	Churn Value	1 = the customer left the company this quarter. 0 = the customer remained with the company. Directly related to Churn Label.	Boolean
3.	Churn Score	A value from 0-100 that is calculated using the predictive tool IBM SPSS Modeler. The model incorporates multiple factors known to cause churn. The higher the score, the more likely the customer will churn.	Integer
4.	CLTV	Customer Lifetime Value. A predicted CLTV is calculated using corporate formulas and existing data. The higher the value, the more valuable the customer. High value customers should be monitored for churn.	Integer
5.	Churn Reason	A customer's specific reason for leaving the company. Directly related to Churn Category.	String

There is also a huge difference between the distribution of the Churn Label Yes (26.5%) and No (73.5%) which is dependent variable for the classification analysis. Hence standardization is required before the training to ensure one feature does not dominate the other.

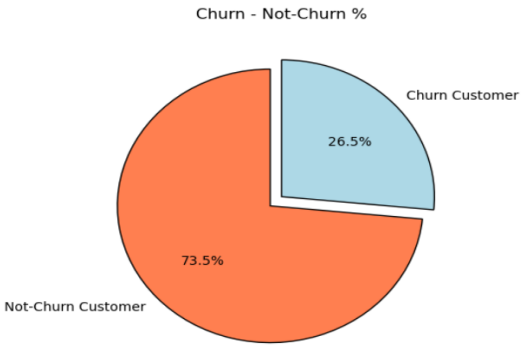

Figure 2: Churn label distribution

5.3 Data Preparation

Data preparation phase involves cleaning the data to handle outliers, missing values and unknown or null values, transforming data into suitable format and encoding categorical values. This step is crucial before building any machine learning model as raw data might give biased or skewed results. Along with this exploratory data analysis (EDA) is also conducted. This pivotal stage aims to enhance the data quality and mitigate any biases that could potentially compromise the analysis. Python libraries Pandas, NumPy, Seaborn and Matplotlib where used.

Exploratory Data Analysis and Data Pre-processing:

During the Data Preparation phase of the CRISP methodology, essential tasks such as data cleaning and exploratory data analysis (EDA) are conducted prior to the modelling phase. This pivotal stage aims to enhance the data quality and mitigate any biases that could potentially compromise the analysis.

During the Exploratory Data Analysis (EDA) phase, Python was utilized to examine the data types of all columns. Most of the columns were found to be represented as objects, indicating the presence of string or categorical data within them.

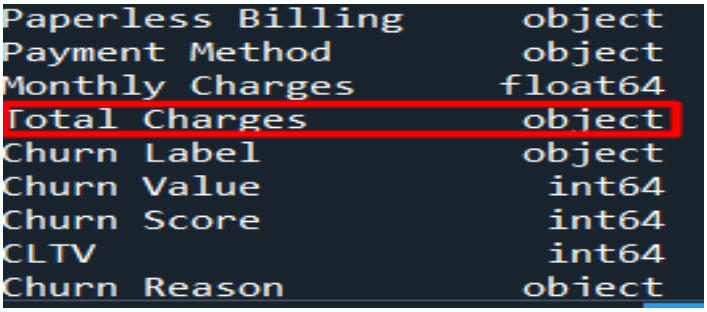

Figure 3: Data type of the attributes.

The Total Charges column stands out as a notable exception within the dataset, as it is represented as an object type rather than a float. This observation implies the possible presence of non-numeric values within the column. Furthermore, upon examining the data frame, it becomes evident that there are 11 instances of null values associated with the Total Charges column. Consequently, the data type for Total Charges was initially set as an object due to the presence of these null values.

Paperless Billing	Payment Meth	Monthly Charges	Total Charges
Yes	Bank transf…	52.55	nan
Yes	Bank transf…	61.9	nan
No	Credit card (automatic)	56.05	nan
No	Mailed check	73.35	nan
No	Mailed check	20.25	nan
No	Mailed check	25.75	nan
No	Mailed check	19.85	nan
No	Mailed check	25.35	nan
No	Mailed check	20	nan
Yes	Mailed check	19.7	nan
No	Mailed check	80.85	nan

Figure 4: Total Charges column with null values

It is essential to change the data type of the Total Charges before proceeding to the next step which is checking missing or null values in all the columns.

```
# Changing the data type of Total Charges from Object to numeric.
telco_df['Total Charges'] = pd.to_numeric(telco_df['Total Charges'], errors='coerce')
```

Figure 5: Data type conversion – Total Charges column.

After the change of the data type, it was discovered that 11 missing values had emerged within the same column.

```
Payment Method      0
Monthly Charges     0
Total Charges      11
Churn Label         0
Churn Value         0
Churn Score         0
CLTV                0
```

Figure 6: Missing values in the Total Charges column

The 11 missing values were imputed with a value of 0 to ensure data completeness and facilitate further analysis.

```
# Replaced missing values with 0
telco_df['Total Charges'].fillna(0, inplace=True)
```

Figure 7: Replacing 0 with missing values – Total Charges column.

Missing values are also found for “Churn Reason” column, it had 5174 missing values, these were replaced with value “Unknown”. Columns “'CustomerID', 'Count', 'Lat Long', 'Zip Code', 'Latitude', 'Longitude' are dropped as they are not significant for the analysis.

Ensuring the integrity of the data set is paramount, and therefore, it is crucial to perform a thorough check for duplicate rows. However, after a meticulous examination, no duplicate rows were detected in the data set. This outcome affirms the consistency and uniqueness of the records within the dataset, contributing to its reliability for subsequent analyses and decision-making processes.

```
In [33]: print(telco_df.duplicated().sum())
0
```

Figure 8: Checking duplicate data values.

Outlier detection performed using z – score method. For each numerical feature z – score is calculated using formula z = (x - mean) / std, where x is value of the feature, mean is the mean value of the feature and std is standard deviation of the feature. Threshold value is set 3, any observation with the z – score greater than 3 is considered as a outlier. There are no outliers found in the data set.

Divided features into two groups named categorical and numerical features for the visualization purpose. Created Histograms of the numerical variables to check their distribution and found that none of them are normally distributed so normalization is required before training the model.

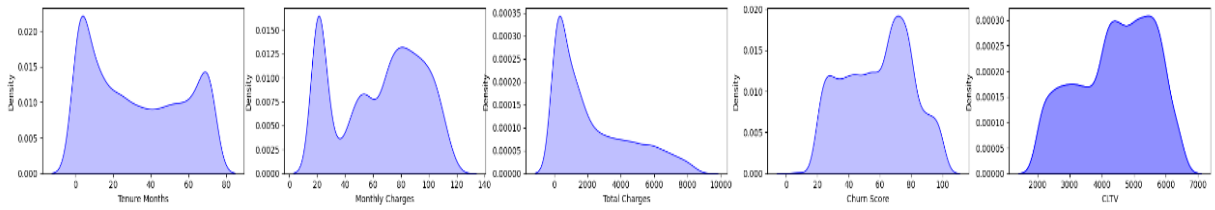

Figure 9: Distribution of numerical variables.

Further two tables are created named churned customers and non – churned customers. It depicts descriptive statistics computed for each subset to understand the difference in various features between the two groups.

It is clear for the above figure that there is a huge difference between the mean value of CLTV, Total Charges and Tenure Months of the churned and non – churned customers.

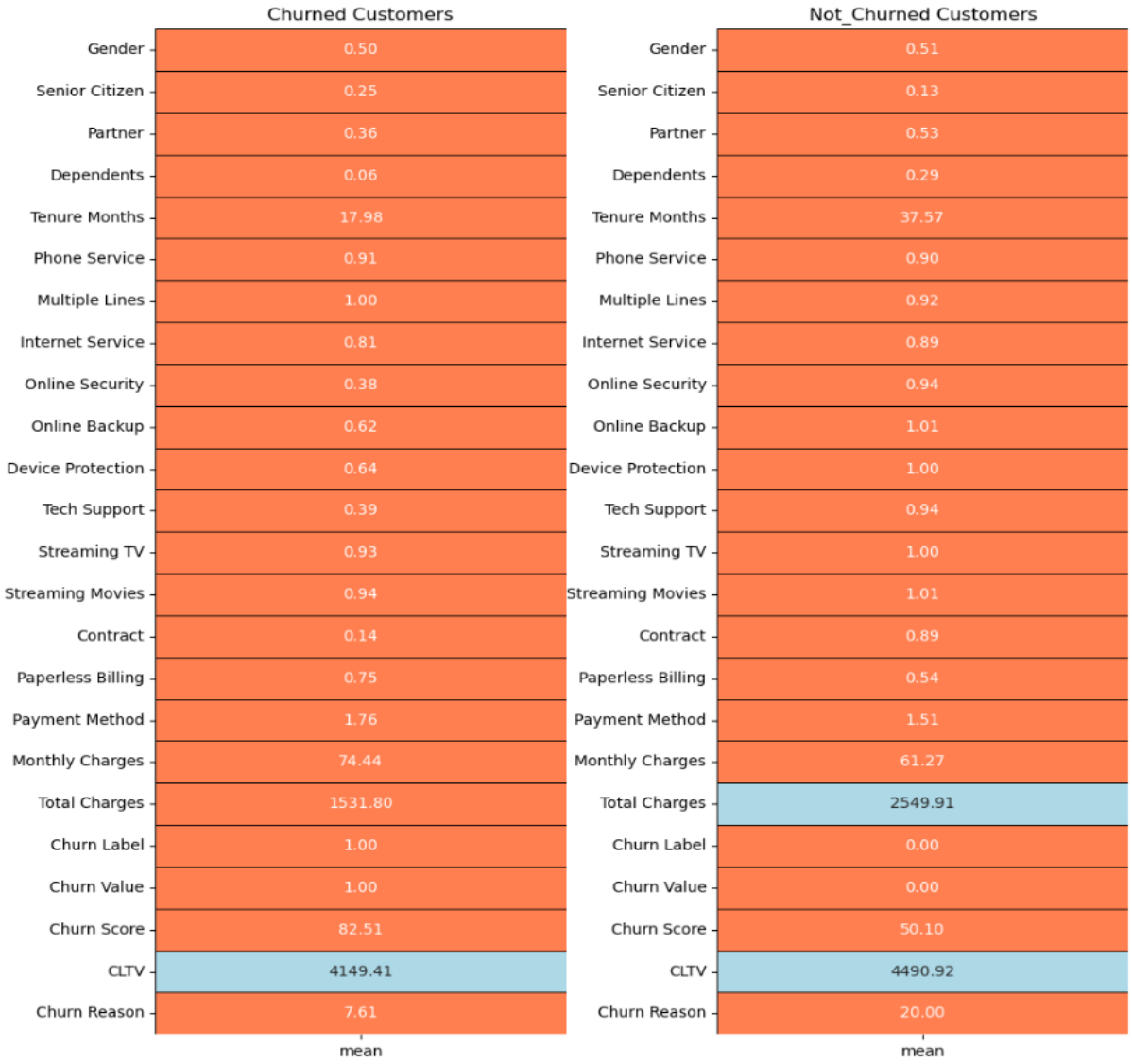

Figure 10: Factors influence on churned and non churned customers.

Data Visualization:

Aim of the data visualization is to represent the data in visual format such as graphs, charts, maps to provide better understanding and interpretation of the data more effectively. Here, Tableau is utilized to create the visualization of chosen data set to address the business problems.

1. **Top factors influencing customer churn.**

   The bar chart presented in Fig 11 provides 20 different reasons for customers to churn with count of customers on Y-axis and churn reasons on x-axis. Starting with support personnel attitude being the primary reason for customer churn, resulting in 192 customers not with the company anymore. The second most leading reason is superior services by competitor such as better data offerings, higher data speed packages, better price offering which ended up combined customer churn count to 621. Remaining factors for churn are in decreasing order which can be of less concern

compared to support and services offered by the company.

By observing the above pattern in the graph, it is clear that company should aim at providing customer relationship training , guidelines and regulations for support personnel to improve their service. Also, the company should enhance and make changes in offerings in various sections of the service in order to gain competitive advantage over other companies and also meet the customer expectations. However, the churn factor such as moved to different countries cannot be addressed by the company as its not possible for the company to develop strategy to prevent churn.

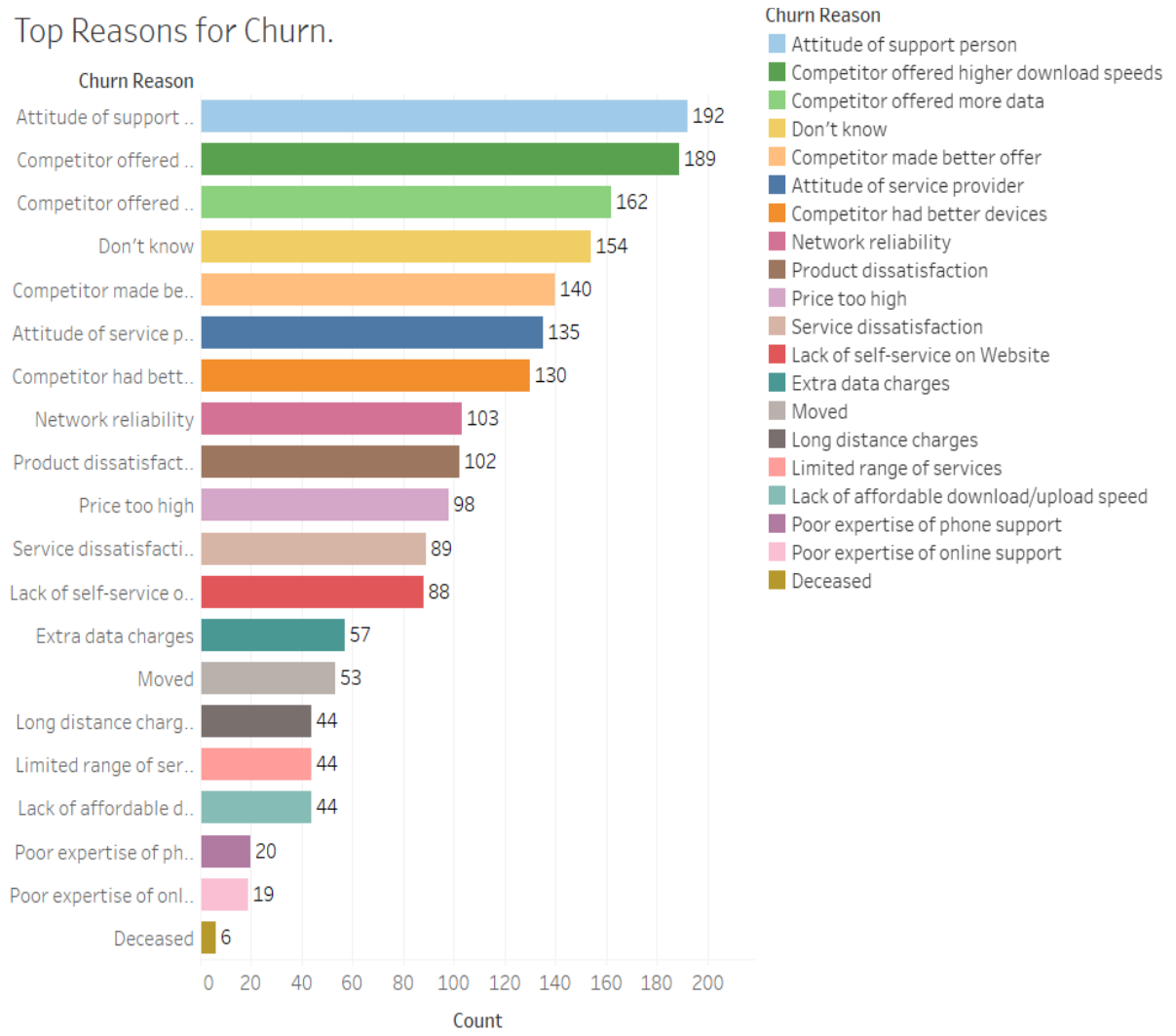

Figure 11: Top Churn Reasons – Bar chart.

2. **Customer churn based on contract type owned by customer.**

The pie chart in figure 11 displays 3 different types of contract (subscription) the company is offering to its customers and their churn rate.

Customers with Month-to-Month contract have a churn rate of 23.5% whereas customers with one year and two-year contract have 2.36% and 0.68% of churn rate respectively. From this visualization it is observed that long term contract customers have very less probability of churn, adding more value to the company. Further

improvement in marketing strategy such providing attractive incentives to customers to convert to long-term subscription, can help the company to increase retention rates.

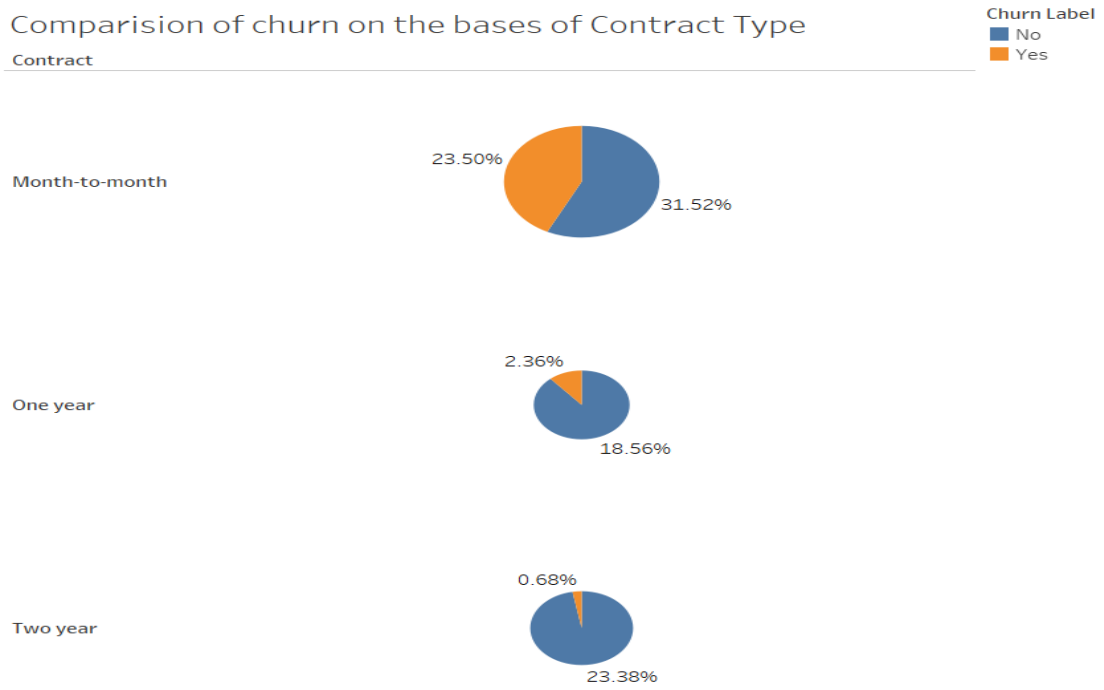

Figure 12: Customer churned based on contract type– Pie chart.

3. **Comparison of customer churn with respect to contract type and gender.**
   The stacked bar chart displays that the churn rate is similar for both female and male customers, irrespective of their chosen payment methods. This indicates gender does not affect customer churn when considering different contract types.

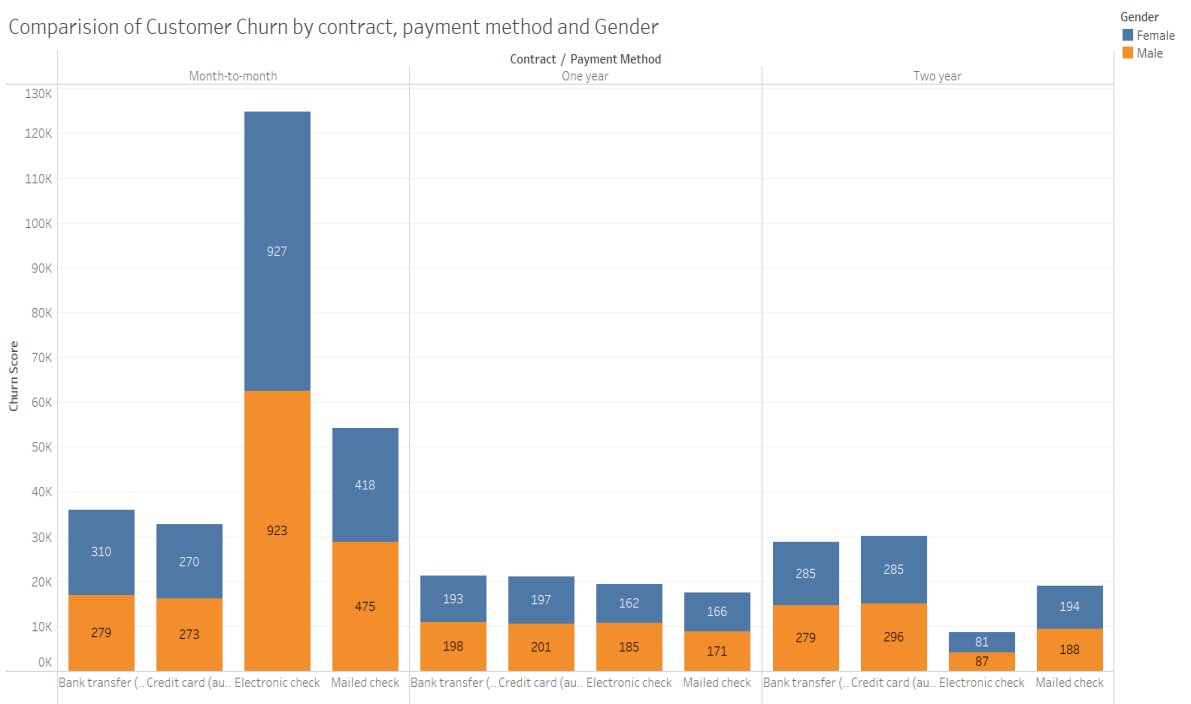

Figure 13: Customer churned based on contract type– stacked bar chart.

4. **Customer Churn behaviour with respect to paid addons.**

   The bar graph in figure 12 displays how paid addons affect customer churn. We can observe the few customers churn for additional paid addons such as back up and security ,device protection plan and premium tech support but there isn't much difference between customers who churned due to high cost with addon and without addon, suggesting that the basic product pricing affects churn more than the addon pricing.

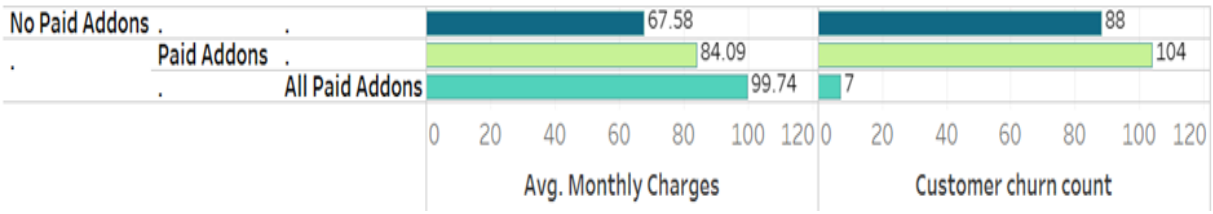

Figure 14: Customer churned based on contract type– Bar chart.

5. **Comparison of Customer Lifetime Value with Customer Churn Count and Analysis of Churn Reasons.**

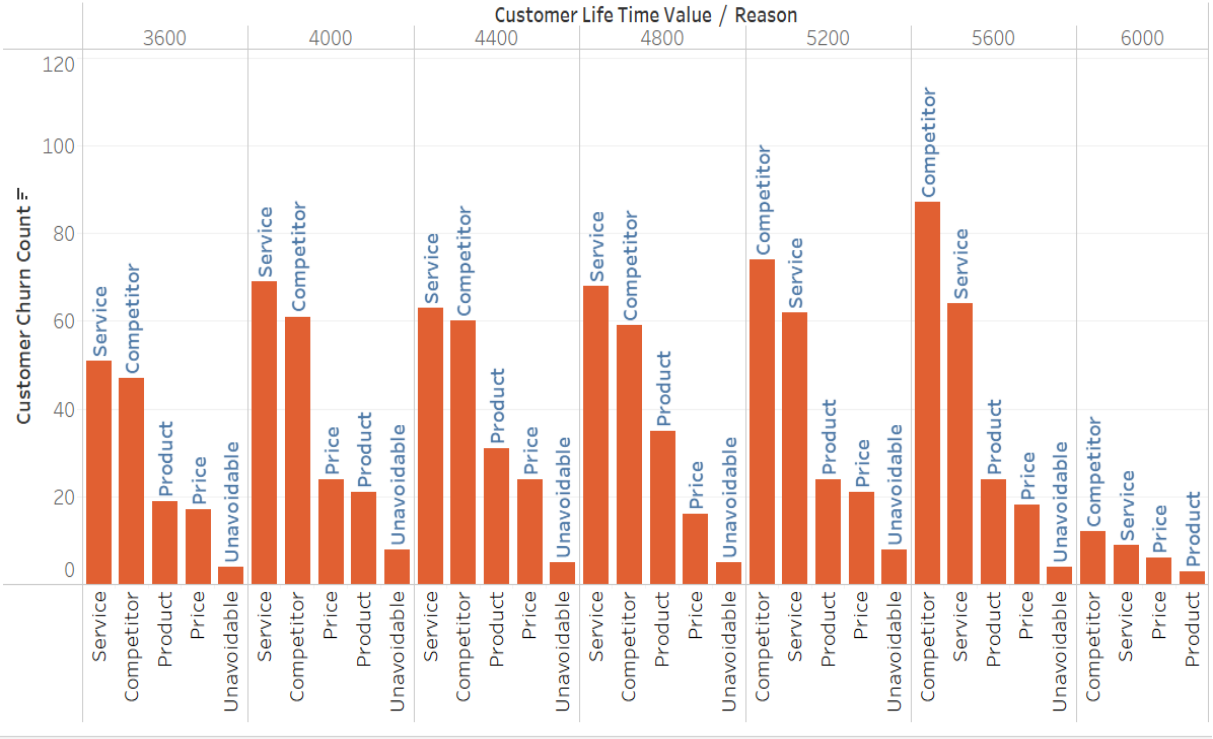

Figure 15: Churn reason comparison with CLTV – Bar chart.

The bar graph presented above illustrates the relationship between Customer Lifetime Value (CLTV), Customer Churn Count, and the reasons for customer churn. CLTV serves as a metric to assess the value of customers to the company, with higher scores indicating greater value. The analysis focuses on customers with intermediate to higher CLTV scores, ranging from 3600 to 6000.

From the graph we can observe that customer churn is primarily driven by the quality of service provided by the company for the customers with CLTV score ranging from 3600 to 4800. Therefore, it is crucial for the company to upgrade various service offerings and improve overall customer satisfaction to reduce the churn rate. Additionally, the company is losing customers to competitors, suggesting the need to devise new and attractive plans and services that can provide a competitive edge in the market.

Furthermore, it is noteworthy that customers in the higher CLTV segment, scoring from 5200 to 6000, express satisfaction with the products and pricing offered by the company. However, churn still occurs due to the competitive advantages of rival firms and suboptimal service levels.

Feature Engineering:

Created a deep copy of the original data set and label encoding the text data. Modification in the original data set will not be highlighted in this copy. Hence, we will use this deep copy that has all the features converted to the numeric values for the visualization and modelling.

Dropped unnecessary features which are 'CustomerID', 'Count', 'Lat Long', 'Zip Code', 'Latitude', 'Longitude', 'Country', 'State', 'City'.

Data Scaling:

Normalization is carried out on columns “Tenure Months”, “Monthly Charges”, “CLTV” and “ Total Charges” using MinMaxScaler as the column value varies significantly when compared to other features. Moreover, standardization is carried out of the categorical variables using StandardScaler.

Correlation:

The correlation matrix heatmap is a valuable data visualization technique utilized to depict the interrelationships among variables within a given dataset. This visualization tool effectively presents the correlation coefficients between pairs of variables in the form of a color-coded matrix. The colour intensity within the heatmap signifies the strength of the correlations, with darker shades indicating stronger associations. By employing the correlation matrix heatmap, analysts can swiftly grasp the extent and nature of relationships existing within the dataset, aiding in the identification of noteworthy patterns, and informing subsequent analytical decisions.

Below is the corelation matrix which shows the correlation between variables. To analyse the correlation matrix of the data set, Python’s Pandas, Matplotlib, Seaborn and NumPy library is used. The heatmap indicates that there are several features that are strongly correlated with each other, while others are weakly or not at all correlated. The Spearman's rank correlation coefficients between the columns of the telco_df data frame is calculated and stored the result in ‘corr_matrix’.

```
plt.figure(figsize=(20,8))
corr_matrix = telco_df.corr('spearman')
sns.heatmap(corr_matrix, cbar=True, annot=True, mask=np.triu(np.ones_like(corr_matrix, dtype=bool)), fmt='.3f', cmap='coolwarm')
```

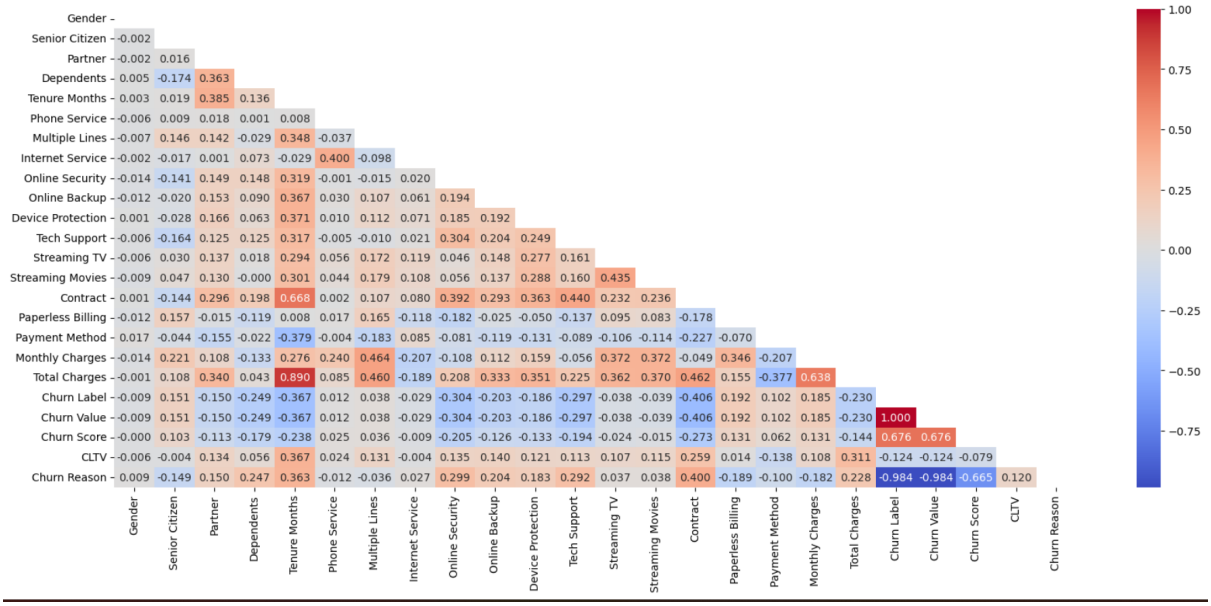

Figure 16: Corelation between variables – Heat map

Moreover, there is a strong positive correlation between the 'Tenure Months' and 'Total Charges'. which indicates that customers who have been with the company for a longer time tend to generate more charges. However, there is a strong negative correlation between the 'Churn Score' and 'Tenure Months', which indicates that the customers who have been with the company for longer time are less likely to churn.

Feature Selection:

Dependent variable is Churn Value. Based on the corelation matrix features with the corelation greater than 0.05 with the dependent variable Churn Value is selected for the modelling and other variables are excluded.

Selected features:

'Senior Citizen', 'Partner', 'Dependents', 'Tenure Months', 'Online Security', 'Online Backup', 'Device Protection', 'Tech Support', 'Contract', 'Paperless Billing', 'Payment Method', 'Monthly Charges', 'Total Charges', 'Churn Score', 'CLTV'.

Non - Selected features:

'Gender', 'Phone Service', 'Multiple Lines', 'Internet Service', 'Streaming TV', 'Streaming Movies'. But features with the less corelation may be affect the model so these variables are not dropped since we have only 37 variables. But as we can see that tenure month is highly correlated with the Total Charges, so Tenure Month is dropped to avoid multicollinearity.

Data Balancing:

To resolve the imbalance in dataset, oversampling is performed on the data. Oversampling increases the minority samples of the target variable churn to match the majority samples. SMOTE (Synthetic Minority Oversampling Technique) is used to create instances of minority class by identifying similarities between the existing minority class samples.

Data set is split into 70% train and 30% test data.

```
import imblearn
from imblearn.over_sampling import SMOTE
from sklearn.model_selection import train_test_split

#model trainning.

X = df1.drop(['Churn Value','Tenure Months','Churn Reason','Churn Label'], axis=1)
y = df1['Churn Value']
# Implement Oversampling (SMOTE) to balance the dataset
X, y = SMOTE(sampling_strategy=1, random_state=0).fit_resample(X, y)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
```

Figure 17: Data splitting – Code snippet

5.4 Modelling

The purpose of this stage is to build a model that can solve business problem or objectives that has been identified in business understanding phase of CRISP-DM. There are different types of modelling such as predictive and classification.

As the objective of our project is to determine whether the customer will churn or stay, the classification model has been chosen in order to identify which category the customer will be segregated. To determine the most suitable classification model for identifying customer churn, a comprehensive evaluation was conducted. Multiple models were employed, and their performance was rigorously assessed, and accuracy of each model has been reported.

Logistic Regression:

As Logistic Regression is well suited for binary classification problems it can used to determine weather or not customer would churn. The model also anticipates the likelihood of churn and then maps the odds into sigmoid curve in order to categorize customers to churners or non churners. The accuracy achieved through logistic regression is 0.93.

```
'Model': 'Logistic Regression',
'Accuracy': 0.9317230273752013,
'Precision': 0.9193846153846154,
'Recall': 0.9485714285714286,
'F1 Score Weightened': 0.9316893567335933,
'AUC ROC': 0.9314752567693744}
```

Decision Tree:

A Decision Tree is a supervised machine learning algorithm that makes decisions using a tree-like structure. Each internal node represents a feature or attribute, each branch represents a decision based on that feature, and each leaf node represents the outcome. By analysing these attributes, the model identifies key performers for contributing the churn. The accuracy achieved through this model is 0.92.

```
'Model': 'Decision Tree',
'Accuracy': 0.9281803542673108,
'Precision': 0.9251572327044025,
'Recall': 0.933968253968254,
'F1 Score Weightened': 0.9281736473221299,
'AUC ROC': 0.9280952380952382}
```

Random Forest:

Random Forest creates an ensemble of decision trees and perform classification by aggregating individual tree predictions. This model considers various customer attributes such as demographics and usage patterns which can predict the probability of churn accurately. Further the model also eliminates the risk of overfitting making it easy to identify costumers at risk. The accuracy achieved through the model is 0.92.

```
'Model': 'Random Forest',
'Accuracy': 0.952012882447665,
'Precision': 0.9384993849938499,
'Recall': 0.9688888888888889,
'F1 Score Weightened': 0.9519884898410906,
'AUC ROC': 0.951764705882353}
```

XGBoost:

XGBoost stands for Extreme gradient Boosting. As the dataset is imbalanced XGBoost has the ability to train the data without performing data balancing task. XGBoost can also learn from churned customer behaviours, determine key indicators which help for accurate predictions of future churners. The accuracy achieved through the model is 0.95.

```
'Model': 'XGboost',
'Accuracy': 0.9500805152979066,
'Precision': 0.9420921544209215,
'Recall': 0.9606349206349206,
'F1 Score Weightened': 0.9500683090429404,
'AUC ROC': 0.9499253034547152}
```

K-Nearest Neighbour:

The K-Nearest Neighbour model can be used to calculate the distance between the target customers and its K nearest neighbours based on factors such as customer subscription pattern, their engagement with company and demographics. By considering the majority group among the nearest neighbours KNN can predict the customer churn probability. The accuracy achieved though this model is 0.80

```
'Model': 'KNN',
'Accuracy': 0.8080515297906602,
'Precision': 0.7569553805774278,
'Recall': 0.9155555555555556,
'F1 Score Weightened': 0.8055512536771907,
'AUC ROC': 0.806470588235294}
```

Neural Network with Sigmoid Function.

For binary classification tasks, a neural network with a sigmoid activation function is commonly used. The logistic function, also known as the sigmoid function, maps the neural network output to a value between 0 and 1, which can be interpreted as a probability or confidence score for the positive class.

```
'Model': 'NN_sigmoid',
'Accuracy': 0.9317230273752013,
'Precision': 0.9267376330619912,
'Recall': 0.9396825396825397,
'F1 Score (Weighted)': 0.931712583761682,
'AUC ROC': 0.9834007677144931}
```

AdaBoost.

AdaBoost follows ensemble learning approach which combines weak classifiers to combine strong predictive model. By adjusting the weights of misclassified attributes the model captures the best patterns in customer behaviour. This adoptive nature of the model help to identify key factors for churn. The accuracy achieved through the model is 0.94.

```
'Model': 'AdaBoost',
'Accuracy': 0.9455716586151369,
'Precision': 0.932349323493235,
'Recall': 0.9625396825396826,
'F1 Score Weightened': 0.9455439918331834,
'AUC ROC': 0.9453221288515407}
```

LGBM:

LGBM is widely used to handle the large data set accurately. It is a open-source gradient boosting framework developed by Microsoft. It is a trustworthy option for reliably and efficiently anticipating customer churn due to its capacity to capture complicated patterns and interactions. The accuracy achieved through the model is 0.95.

```
'Model': 'LGBMClassifier',
'Accuracy': 0.9526570048309179,
'Precision': 0.9412855377008653,
'Recall': 0.966984126984127,
'F1 Score Weightened': 0.952638408270336,
'AUC ROC': 0.9524463118580765}
```

CatBoost.

CatBoost is mainly designed to handle categorical variables in machine learning tasks and provides powerful capabilities to handle categorical features without encoding or pre-processing. The accuracy achieved through the model is 0.954.

```
'Model': 'CatBoostClassifier',
'Accuracy': 0.9539452495974236,
'Precision': 0.9441687344913151,
'Recall': 0.9663492063492064,
'F1 Score Weightened': 0.9539307461982957,
'AUC ROC': 0.9537628384687208}
```

Support Vector Machine (SVM).

SVM is a powerful supervised learning algorithm that can be used for both classification and regression tasks. SVM seeks the best decision boundary or hyperplane that separates data points from different classes while maximizing the margin between them. The accuracy achieved through the model is 0.92.

```
'Model': 'SVM',
'Accuracy': 0.9265700483091788,
'Precision': 0.9045045045045045,
'Recall': 0.9561904761904761,
'F1 Score Weightened': 0.9264773337237104,
'AUC ROC': 0.9261344537815127}
```

5.5 Evaluation.

ROC (Receiver Operating Characteristics)

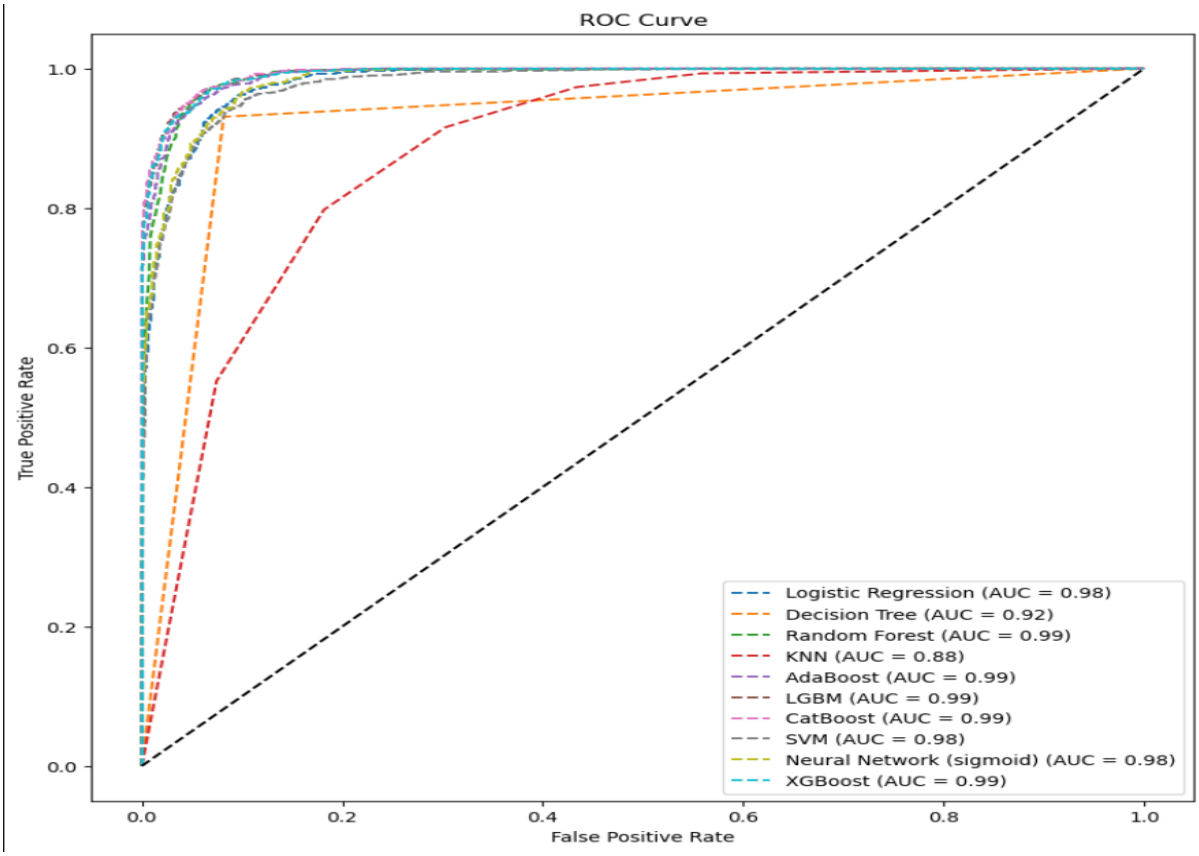

Figure 18: ROC Curve & positive Rate for all models

Comparison of various machine learning model is done using Receiver Operating Characteristic (ROC) and the Area Under the Curve (AUC) metric. We used Scikit – learn library for the analysis and ten models were selected including Logistic Regression, Decision Tree, Random Forest, KNN, AdaBoost, LGBM, CatBoost, SVM, Neural Network (sigmoid), and XGBoost. Every model was fitted to training data set and ROC curve and AUC were calculated using testing data set.

The False positive Rate and True Positive Rate is Displayed on the x and y axis respectively. The models are represented by various coloured lines and labels indicates the model name and AUC score. The models XGBoost, CatBoost and LGBM, AdaBoost and Random Forest achieved the highest AUC score (0.99). The ROC curve of the Top performers are located close to the upper – top left corner which indicates higher TPR for a lower FPR.

This analysis provides valuable insights to compare performance of the various machine learning models which helps in the selection of the suitable model according to the various applications.

Cross Validation.

The primary goal of cross-validation is to predict how well a model will generalize to new, previously unseen data. Cross-validation reduces the impact of randomness and potential bias in a single train-test split by evaluating the model on multiple different subsets of the data. It provides a more reliable evaluation of the model's performance, assisting in the avoidance of overfitting and underfitting issues. Stratified K – Fold Cross – Validation is used to calculate evaluation metrics for the classification models.

	Model	Mean Test Accuracy	Mean Test F1 Score (Weighted)	Mean Test Precision	Mean Test Recall	Mean Test AUC-ROC	Mean Test F1 Score
0	Logistic Regression	0.93	0.93	0.91	0.94	0.98	0.93
1	Decision Tree	0.93	0.93	0.92	0.93	0.93	0.93
2	Random Forest	0.95	0.95	0.94	0.97	0.99	0.95
3	KNN	0.81	0.81	0.76	0.92	0.89	0.83
4	AdaBoost	0.94	0.94	0.93	0.96	0.99	0.94
5	LGBM	0.95	0.95	0.94	0.96	0.99	0.95
6	CatBoost	0.95	0.95	0.94	0.96	0.99	0.95
7	SVM	0.92	0.92	0.90	0.96	0.98	0.93
8	NN Sigmoid	0.92	0.92	0.91	0.94	0.97	0.92
9	XGBoost	0.95	0.95	0.94	0.96	0.99	0.95

Figure 19: Best models F1 Scores

Above table depicts that catboost, XGBoost, LGBM and Random Forest are the best models according to their F1 score of 0.95.

Hyperparameter tunning.

For hyperparameter tuning and performance evaluation, Python's scikit-learn library was used. Each model has a predefined hyperparameter grid that specifies the range of values for various hyperparameters. Accuracy, weighted F1 score, precision, recall, AUC-ROC, and F1 score were the scoring metrics used for evaluation. The best 4 models and those hyperparameters are shown in below figure.

Model	Hyperparameters	Mean Test Accuracy	Mean Test F1 Score (Weighted)	Mean Test Precision	Mean Test Recall	Mean Test AUC-ROC	Mean Test F1 Score
CatBoost	{'l2_leaf_reg': 1, 'learning_rate': 0.1, 'max_depth': 6, 'n_estimators': 200, 'random_seed': 42, 'subsample': 0.9}	0.95	0.95	0.95	0.95	0.99	0.95
Random Forest	{'bootstrap': False, 'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 'log2', 'max_leaf_nodes': None, 'max_samples': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 50, 'n_jobs': None, 'oob_score': False, 'random_state': None, 'verbose': 0, 'warm_start': False}	0.95	0.95	0.95	0.95	0.99	0.95
LGBM	{'boosting_type': 'gbdt', 'class_weight': None, 'colsample_bytree': 1.0, 'importance_type': 'split', 'learning_rate': 0.1, 'max_depth': -1, 'min_child_samples': 20, 'min_child_weight': 0.001, 'min_split_gain': 0.0, 'n_estimators': 100, 'n_jobs': -1, 'num_leaves': 31, 'objective': None, 'random_state': None, 'reg_alpha': 0.0, 'reg_lambda': 0.0, 'silent': 'warn', 'subsample': 1.0, 'subsample_for_bin': 200000, 'subsample_freq': 0}	0.95	0.95	0.95	0.95	0.99	0.95
XGBoost	{'objective': 'binary:logistic', 'use_label_encoder': None, 'base_score': None, 'booster': None, 'callbacks': None, 'colsample_bylevel': None, 'colsample_bynode': None, 'colsample_bytree': None, 'early_stopping_rounds': None, 'enable_categorical': False, 'eval_metric': None, 'feature_types': None, 'gamma': None, 'gpu_id': None, 'grow_policy': None, 'importance_type': None, 'interaction_constraints': None, 'learning_rate': None, 'max_bin': None, 'max_cat_threshold': None, 'max_cat_to_onehot': None, 'max_delta_step': None, 'max_depth': None, 'max_leaves': None, 'min_child_weight': None, 'missing': nan, 'monotone_constraints': None, 'n_estimators': 100, 'n_jobs': None, 'num_parallel_tree': None, 'predictor': None, 'random_state': None, 'reg_alpha': None, 'reg_lambda': None, 'sampling_method': None, 'scale_pos_weight': None, 'subsample': None, 'tree_method': None, 'validate_parameters': None, 'verbosity': None}	0.95	0.95	0.95	0.95	0.99	0.95

Figure 20: Best models in regards to hyperparameter

Confusion matrix.

Below is the confusion matrix for the catBoost model. It is clear that true positive values are 21439 and true negative values are 1521 and false positive and False negative values are 54 and 13.

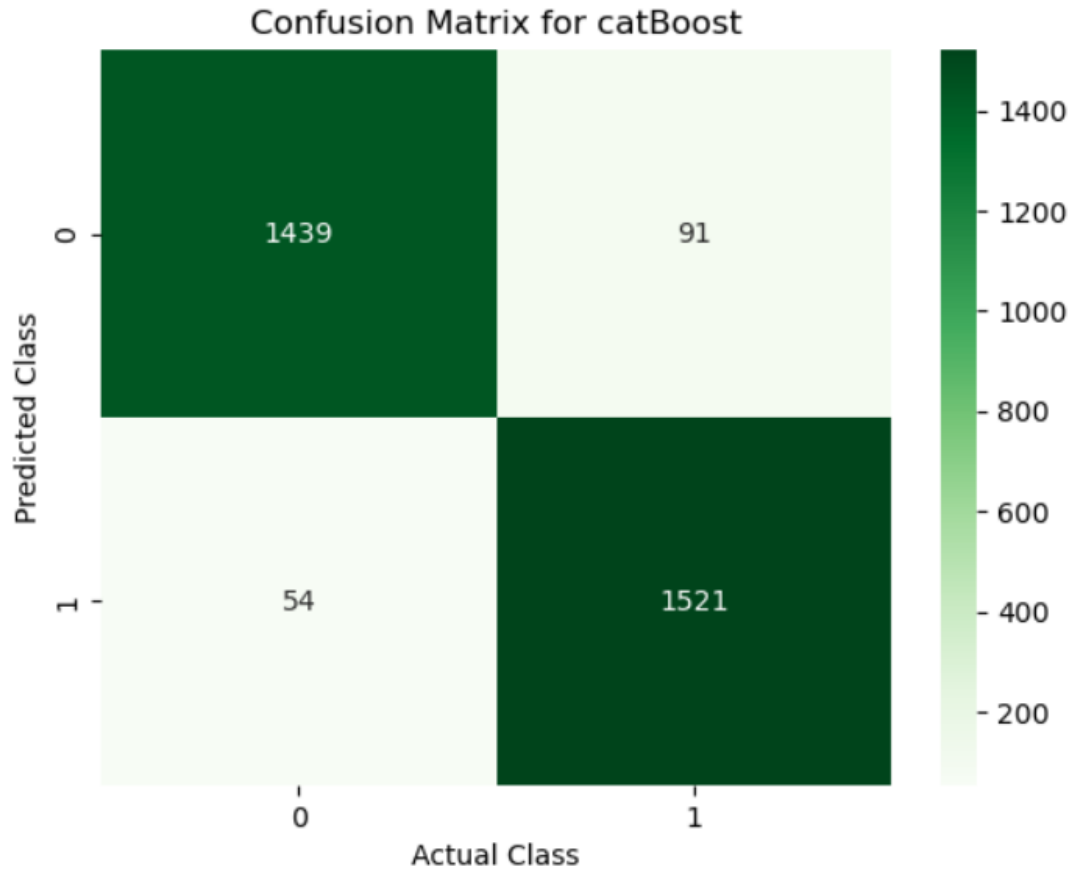

Figure 21: Confusion matrix for the catBoost

In other words, there are 1439 values are 0 and these values are correctly classified by the model. There are 1521 values are 1 and these are also correctly classified by the model. However, there are 91 values are 1 but wrongly classified as 0 by the model and 54 values as 0 but wrongly classified as 1 by the model.

Identifying the feature important from each model.

In below figure, the most important feature affecting the customer churn are identified.

	Unique Features
0	Online Security
1	Monthly Charges
2	Payment Method
3	Total Charges
4	Churn Score
5	Contract

Figure 22: Top Churn reason

6. Sentimental Analysis.

Sentimental analysis also known as opinion mining is used to determine the attitude or emotional state of the customers regarding the service. It involves analysing reviews of the customers and determines the overall expression in the text by using machine learning algorithm to classify the text into positive, neutral, negative category based on the context and tone of the text. Sentimental analysis could be a powerful tool for the company to understand customer's emotions and preventing them from churn. It is performed using VADER sentimental analysis library in Python.

From NLTK library sentimental analyser is initiated using SentimentIntensityAnalyzer() function, then sentiment scores for each review is computed using polarity_scores() function and extracted the compound sentiment score which ranges from -1 to 1. Each review is classified as positive or negative based on the compound score.

Result shows that out of 7043 reviews 92.8% reviews indicates the positive sentiments about the company and 7.2% reviews are negative. This gave an insight that around 93% customers are satisfied or neutral and 7% customers have completely negative perspective about one or more service provided by the company.

```
# Initialize the sentiment analyzer
sid = SentimentIntensityAnalyzer()

# Compute the sentiment scores for each review
telco_df['sentiment_scores'] = telco_df['Churn Reason'].apply(lambda x: sid.polarity_scores(x))
#print(telco_df['sentiment_scores'])

# Extract the compound sentiment score
telco_df['sentiment'] = telco_df['sentiment_scores'].apply(lambda x: 'positive' if x['compound']>=0 else 'negative')

# Print the percentage of positive and negative reviews
print("Percentage of Positive Reviews: {:.2f}%".format((telco_df['sentiment'] == 'positive').mean()*100))
print("Percentage of Negative Reviews: {:.2f}%".format((telco_df['sentiment'] == 'negative').mean()*100))
```

Figure 23: Sentiment analysis script

```
Percentage of Positive Reviews: 92.80%
Percentage of Negative Reviews: 7.20%
```

Figure 24: Sentiment analysis script

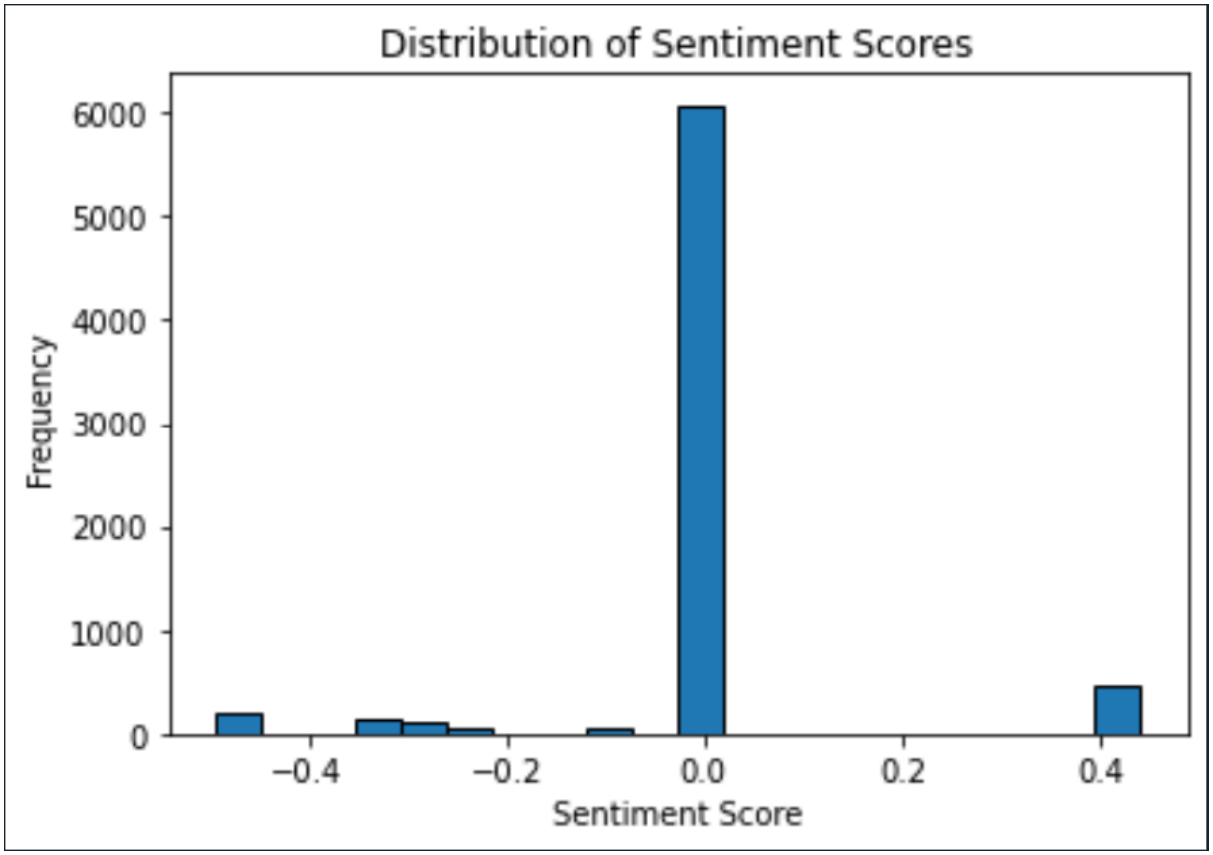

Figure 25: Distribution of Sentiment Score

6.1 Histogram Analysis:

The histogram displays the distribution of sentiment scores for customer churn reasons. The sentiment scores range from -1 (most negative) to +1 (most positive). The histogram shows that the sentiment scores are approximately normally distributed around 0, with a slightly positive skew. Most of the sentiment scores fall in the range of -0.5 to 0.5, indicating that most of the reviews were neutral or had a slightly positive or negative sentiment.

The histogram also reveals that there were a few reviews with extremely negative sentiment scores below -0.4, indicating that some customers had very negative experiences with the telecommunication company. On the other hand, there were also a few reviews with extremely positive sentiment scores above 0.4, indicating that some customers had very positive experiences with the company.

The histogram provides valuable insights into the sentiment of customer churn reasons for a telecommunication company. The analysis revealed that most of the customer churn reasons had a neutral or slightly positive or negative sentiment. However, there were a few customers who had extremely negative or positive experiences with the company, indicating that there is still room for improvement in certain areas of the company's customer service. Top of Form

6.2 Identifying the factors causing negative sentiments.

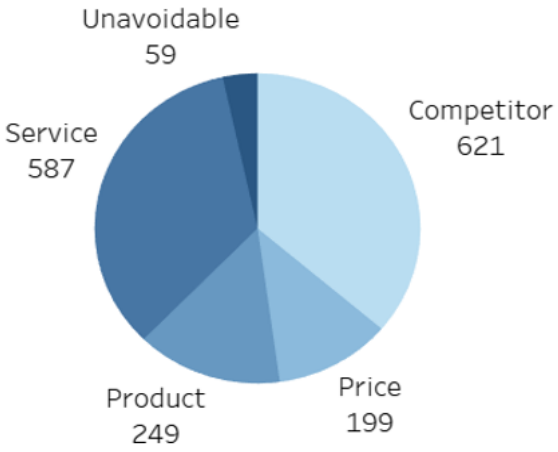

Figure 26: Top Churn reason

To understand the major factors causing negative sentiments, the customer feedback is segmented into 6 clusters, Unavoidable, Unknown, Competitor, Service, Product and Price. The pie chart presented in the figure 10 shows the proportion of these factors. It is evident that competitors and service section of the company are major reason for customers to churn which also corelates with results of graph presented in Fig 9. Observing this pattern, focusing on service segment can help increase the customer satisfaction and retain them. Also, its equally important to devise better strategies to surpass the competition and stay relevant in market. Customer survey or feedback can be conducted to adapt to the changing market conditions and growing needs of the customer.

6.3 Comparison with Corelation results:

From the corelation function it is also observed that apart from the factors such as service and competitors affecting churn, Monthly charges and paperless billing are also contributing to

churn on which the company can develop strategy such as providing customized monthly plan to customers and obtaining periodic feedback on services provided by the company

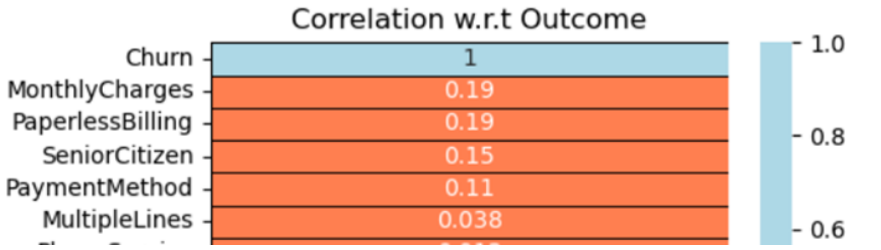

Figure 27: Top Churn reason

6.4 Tools and Technologies used in the Project:

The following tools and technologies were utilized to work on this project, explore the dataset, and derive important insights. Justifying the project's results will depend on this.

- Python
- Tableau
- Microsoft Excel

7. Result and Conclusion

Customer turnover is influenced by several variables, such as pricing, the availability of add-ons, the attitude of support staff, the competitors offering superior services, contract type and payment method.

Findings from the research show that customer attrition is significantly influenced by the attitude of support staff, highlighting the need of putting rules and training programs in place that improve customer service interactions. A major factor in the high rate of customer turnover is also the availability of aggressive competition, which offers innovative data, quicker download speeds, and extended gadget capabilities. The business should focus its efforts on improving its services, taking an inclusive stance, and embracing creative tactics to fulfil a wide range of client expectations to handle this difficulty.

Customers with short-term contracts, such month-to-month subscriptions, are more likely to churn than those with long-term contracts, according to analyses of contract types. The business must develop plans to get clients to sign long-term contracts by touting their benefits and providing alluring incentives.

Majority of clients are happy or indifferent, according to the sentiment analysis of customer evaluations, however a tiny minority express unfavourable opinion about some services. This underlines the need of steadily raising service standards to satisfy clients and reduce turnover.

To model and forecast churn, classifiers from XGBoost, LightGBM, and Random Forest were used. XGBoost achieved a ROC_AUC score of 94.99% whereas each model produced different results.

It is essential for telecom businesses to focus personalisation and improve the entire customer experience to reduce customer attrition. Telecom companies can create specialized plans, services, and marketing promotions by utilizing the data and preferences of their customers. Customer loyalty may be promoted, and churn rates successfully decreased by allocating resources to provide seamless customer experiences across different engagement points, such as mobile applications, websites, and customer care.
Telecom firms may use proactive client retention methods rather than only relying on remedial actions. Continuous monitoring of customer happiness, early detection of causes for discontent, and taking preventative measures to resolve customer problems before they result in churn are all required for this.

A further crucial factor will be value-added services. Offering options other than the standard phone and data plans allows telecom companies to stand out from the competition. The Internet of Things (IoT), smart home solutions, multimedia streaming, and service bundles that target user groups might all fall under this category. Telecommunications firms may improve customer loyalty and lower churn by delivering distinctive and pertinent offers.

Customer satisfaction and retention rates will continue to be significantly impacted by the Caliber and dependability of network services. To increase coverage and guarantee reliable access, telecom businesses must invest in improving their network infrastructure. This calls for enhancing 5G networks to enable quicker data transmission, lower latency, and build more reliable networks.

Overall, the report recommends that the telecommunications business concentrate on upgrading service offerings, improving customer service contacts, and creating customer-centric initiatives to lower churn. The organization may try to retain clients, boost customer happiness, and eventually improve its market position by addressing the identified concerns and utilizing innovative modelling tools.

8. Future Scope

The extent of telecom customer turnover will alter as technology develops, as customers' expectations change, and as the telecoms sector changes. Predictive analytics and artificial intelligence will be one area of emphasis. Advanced analytics methods and AI (Artificial Intelligence) may be used by telecom businesses to spot trends and foretell client turnover. To retain at-risk consumers, providers might take preventative action by examining client data, usage trends, and behaviour. Apart from these techniques such as active learning and survival analysis can be used to increase the efficiency of Churn prediction.

In all, technical improvements, customer-focused initiatives, and the capacity to adjust to shifting market dynamics will all play a role in how much customer attrition there is in the telecom business. Telecom operators may lower customer churn and succeed in a fiercely competitive industry by embracing innovation and placing their customers' needs first.

9. References

1. Agrawal, S., Das, A., Gaikwad, A. and Dhage, S. (2018). *Customer Churn Prediction Modelling Based on Behavioural Patterns Analysis using Deep Learning*. [online] IEEE Xplore. doi:https://doi.org/10.1109/ICSCEE.2018.8538420.

2. Momin, S., Tanuj Bohra and Raut, P. (2020). Prediction of Customer Churn Using Machine Learning. *EAI/Springer Innovations in Communication and Computing*, pp.203–212. doi:https://doi.org/10.1007/978-3-030-19562-5_20.

3. Pamina, J., Raja, B., SathyaBama, S., S, S., Sruthi, M.S., S, K., V J, A. and G, P. (2019). *An Effective Classifier for Predicting Churn in Telecommunication*. [online] papers.ssrn.com. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3399937 [Accessed 17 May 2021].

4. Halibas, A.S., Cherian Matthew, A., Pillai, I.G., Harold Reazol, J., Delvo, E.G. and Bonachita Reazol, L. (2019). *Determining the Intervening Effects of Exploratory Data Analysis and Feature Engineering in Telecoms Customer Churn Modelling*. [online] IEEE Xplore. doi:https://doi.org/10.1109/ICBDSC.2019.8645578

5. Mitkees, I.M.M., Badr, S.M. and ElSeddawy, A.I.B. (2017). *Customer churn prediction model using data mining techniques*. [online] IEEE Xplore. doi:https://doi.org/10.1109/ICENCO.2017.8289798.