SCHOOL OF COMPUTING, ENGINEERING AND DIGITAL TECHNOLOGIES

MIDDLESBROUGH, TS1 3BA

FORECASTING TELECOM INDUSTRY CUSTOMER ATTRITION USING MACHINE LEARNING TECHNIQUES

CIS4055-N-FJ1-2022 - COMPUTING MASTERS PROJECT

Submitted in partial requirements for the degree of

MSc - Applied Artificial Intelligence with Advanced Practice

May 9th, 2023

Supervisor

THAKOR VISHALKUMAR

Submitted by

GOWTHAMI NAGAPPAN (B1326237)

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

TABLE OF CONTENTS

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

ACKNOWLEDGEMENT

My sincere efforts have led to the completion of the project Titled “Forecasting Telecom industry customer attrition using machine learning methods”. I would like to sincerely thank my Supervisor, Thakor VishalKumar who had given his immense support and continuous feedback through the completion of the project. His level of knowledge has been instrumental in providing constructive insights at every stage of the thesis. The ideas imparted in the project were refined at several stages of the thesis through valuable feedback by colleagues which pointed me to disseminate the knowledge gained in a streamlined manner and that has improved the quality of the project. The confidence level and emotional support imparted by all the stakeholders led me to complete the project and their support was imminent throughout the project cycle. Again, I would like to appreciate and thank my supervisor, friends and family for providing guidance and emotional support which helped to complete the project successfully.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

DECLARATION

I declare that the project has been done entirely from my end. From collecting the dataset with the information on Telecom industry services and using this dataset to predict the result of the customer attrition has been completed by myself and I can assure that all the contents of the project is my unique work and does not involve any kind of citation from other academic submissions. In the bibliography, all sources have been appropriately acknowledged. Any support I received in my research and study writing has been recognized. Without consent, I have not used any data, information or material gained from other sources. I have not permitted and will not allow anyone to replicate my work and pass it off as their own. I am aware that any violation of academic integrity or dishonesty might have serious consequences.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

FORECASTING TELECOM INDUSTRY CUSTOMER ATTRITION USING MACHINE LEARNING TECHNIQUES

ABSTRACT

The customers are always considered as the kings of Business. Telecom firms consider their consumers as a Lifetime Value which helps them in calculating the profits for a certain period. Customer Lifetime value is a powerful and straightforward indicator which combines the profitability and attrition risk level of each customer. To maximize the profits, businesses can use customer lifetime value to provide various loyalty programs and retention care plans for their existing customers. Customer Lifetime Value also helps the business to plan on, how to provide the right services for new consumers. Nowadays, the Telecom industries are facing lots of difficulties in maintaining the customers. Customers Turnover and higher rate of attrition have impacted their growth and profitability which in turn makes the industry to run in a loss. This project attempts to construct a predictive model using machine learning techniques which helps the industries to reduce the risk of attrition and the ways to retain the existing customers. This project utilized the IBM sample dataset which contains customer demographics, usage habits and service history. Multiple methods like logistic regression, random forests, k-neighbors classifier and others machine learning models have been implemented in this project to anticipate customer attrition levels. Then the models are evaluated to check the performance through accuracy, precision, and recall. Customers monthly service costs and the tenure period of the customers are the two main variables used in Telecommunication industry to calculate the Customer Lifetime Value. Customer monthly profits are obtained from accounting models and it is mandatory to analyze the customer survival graph using specific parameters which helps in predicting the customer Lifetime Value. To improve the business and brand value, customer lifetime value is calculated using various analysis and methods and suggestions are made to improve the service. The most important factors of attrition according through this project analysis are consumer usage, billing type, contract period, monthly and total charges with add on services such as TV streaming, internet, and other services. These factors helped to create a machine learning model which assisted in forecasting the rate of customer attrition and had an accuracy of 92%. Based on the findings of Explanatory Data Analysis, client retention methods such as targeted promotions, specialized service plans and pro-active customer support have been suggested. This approach and suggestions will help telecom firms by lowering customer turnover and improving loyalty and pleasure.

Eventually this study shows how machine learning methods can be possibly used in web applications to forecast telecom customer attrition and offer tactics to retain the customers in Telecommunication Industries.

Keywords: Machine Learning, Customer retention, Customer Turnover, Customer Attritions, Explanatory Data analysis, Telecom industries, Logistic regression, Decision trees, Targeted promotions, Care plans for existing customers.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

INTRODUCTION

The Telecommunication sector is a competitive sector in the industry. Every telecom industry differs from each other by offering cutting edge services and by providing alluring deals along with top notch care for their customers. Even though all businesses are doing their best to stand high in the crowd but they still face the high rate of customer attrition which impacts the business and makes industry stand a step behind. A recent survey has found that the average annual turnover rate for US telecom industry employees is around 24% and they have also stated that some businesses face even more rates than this [1]. High attrition rates will always have an impact on profitability, growth and reputation which leads to a significant problem for business and thus leads to a higher rate of competition.

In the current era, customers habits and preferences do not remain the same and the customers are continuously changing based upon the needs and satisfaction with the services from the industry. The telecommunications industry should also work in the same way to satisfy and increase the growth of the customers. According to recent research by "Grand View Research" anticipated that telecommunication market will grow more than 6% from now by the year 2030. The growth will be induced by several factors such as rising internet adoption, demand for wireless communication and the development of emerging technologies like 5G and the Internet of Things (IoT) [2].

This Project explains in brief the investigations made earlier into customer attrition and customer retention tactics in the telecom industries. For Example, Kavitha et al (2019) presented a study on the "Customer Churn prediction in the Telecom industry" using methods which are decision tree, random forest and logistic regression algorithms. This experiment was on a real-world dataset and found that the random forest algorithm performed better than other algorithms in the Realtime dataset [3]. Additionally, Ahmad et al (2020) Conducted research on customer leftovers in Telecom industry using ML algorithm in a big data platform. This experiment used several algorithms such as decision tree, logistic regression algorithm and SVM - Support Vector Machine to predict customer attrition. This study also analyzed the impact of the unique features on customer churn and found that the billing information was the most crucial for prediction of customer churn [4].

Nowadays many companies are using machine learning and sophisticated techniques to know the customers who are with the intention of leaving and the companies are taking preventive measures to retain the customers. By using these tools, businesses can now identify which customers are about to leave and the firm can plan accordingly to retain the customers.

With the help of machine learning techniques, the company can now find clients who have the intention to leave the telecom sector and offer suggestions to keep customers. To evaluate various machine learning models and decide the most suitable model (based on performance metrics) with the dataset including customer demographics, usage habits and service history.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

In the Telecommunication industry, factors like usage patterns, service plans and consumer demographics are the key factors of attrition prediction. Then, these measures are taken to develop a model which helps to find the clients who are about to leave. Companies can offer retention plans such as targeted discounts, personalized service plans and initiative-taking customer care which can aid lower customer attrition and improve customer satisfaction and loyalty based on the findings of investigation.

PROJECT AIM

The purpose of this research is to employ machine learning approaches to create a best model for customer attrition in the telecom industry. This model will examine data on user demographics, usage habits and service history which are the main factors that lead to a higher rate of customer attrition. Based on this research, the model is created to supply retention plans which helps the telecom industry to keep the existing customers and helps the sector to satisfy the needs, wants of the customer and improve their happiness which in turn leads to lower level of customer attrition rates.

PROJECT OBJECTIVE

The importance of this research comes with the intention to offer the telecom sector a data driven strategy to reduce customer attrition and increase the rate of customer retention. Through machine learning techniques, companies can get insights about the customers behaviors and preferences which helps to have a focus on the needs of the customers and it also helps to have a tactic to remain the customers.

The goal of the project is to show how machine learning can be used to solve the problems related to customer attrition in the telecom sector, improve customer satisfaction and promote corporate growth. The main aim of this project is to create a robust predictive model for client attrition in the telecom sector by briefing the findings of research and using innovative machine learning techniques. As an outcome of the findings, the telecom sector will be able to develop focused and individualized retention strategies that will increase customer happiness.

The following goals will be pursued to reach the goal:

- Develop and process data on customer demographics, usage habits and service history from the telecom industry.
- Analyzing the data using machine learning techniques reveals major factors that play a key role in client attrition.
- Design a predictive model using supervised ML methods like logistic regression, decision trees algorithm, support vector machines and other models to predict customer attrition.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

- Examine the predictive model performance using suitable metrics which are accuracy, precision, recall and F1 score.
- To develop targeted retention strategies based on the predictive model's observations.

RESEARCH QUESTION

How machine learning technique is used to examine customer attrition in the telecommunication business and what kind of targeted retention strategies can be developed based upon the predictions?

This project will collect and examine data from consumer demographics, usage trends and service history from telecommunication industry. Numerous machine learning algorithms are applied to analyze the data which contributes as a major factor for client attrition. Based on the outcome, the predictive model will be assessed and evaluated using relevant metrics and targeted retention tactics will be proven.

This project provides insights about the applications of machine learning techniques which help to reduce the customer attrition rate and improve the loyalty and satisfaction of the customers. The findings of this project will enhance the growth of the telecom sector which boosts growth and competition in the market.

PROJECT STRUCTURE

Data Collection: This is the first step of the project. The Data was collected from publicly available IBM sample Telecom Fictional dataset which has the consumer demographics, usage habits and service history.

Data Preprocessing: The second step is data preprocessing. In this step the data is cleaned and transferred the collected data into an analysis-ready-format. Data cleaning, feature selection and data normalization will be needed to do this.

Exploratory Data Analysis: To gain insights about the dataset and find the significant elements that play a significant role in customer attrition, the preprocessed data will be studied through exploratory data analysis.

Model Development: This is the decisive step in the project. To establish a predictive model for customer attrition using supervised ML methods like logistic regression, decision trees algorithm, support vector machines and other models. Based on the preprocessed data, the model will be trained and assessed using relevant metrics like accuracy, precision, recall and F1 score.

Retention Strategy Development: Targeted retention Strategies will be developed based upon the observations of predictive model. To increase the customer satisfaction and retention rates, these techniques will be adapted according to individual clients demands.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

LITERATURE REVIEW

The telecommunications industry is the most competitive sector in the market with major customer attrition rates. A range of factors which induce customer attrition rates have been discussed above which includes age, geography, billing problems and call waiting etc. In the most recent years, Telecom industries have started to use machine learning and interest has been increasing till now with no doubts. As said, machine learning plays a vital role in predicting customer attrition and it helps the telecom sector to produce various techniques and models to retain the customers.

Vafeiadis et al (2019) compared several Machine learning techniques which are decision trees algorithm, SVM - Support Vector Machine and Neutral Networks to predict the Customer turnover in the Telecom industry. They used real time dataset for this experiment and concluded that SVM outperformed other algorithms in customer churn prediction [5]. In the paper by Ammara Ahmed et al, it provides a comprehensive review and analysis of the churn customer prediction by classifying these methods into three categories such as statistical method, machine learning and hybrid methods. This paper explains the advantages and disadvantages of each model by comparing each model's performance [6]. Another paper published by Praveen Lalwani et al with six machine learning algorithms including artificial neural network emphasizes the importance and concluded the gradient boosting has higher performance over other models [7]. Overall, these papers give a valuable insight into the current state of art in attrition prediction methods and developing effective churn prediction system for the researcher.

Sahar F sabbeh et al compares the performance of three machine learning algorithms like as decision tree, artificial neural network method and support vector machine. The performance of decision tree algorithm was good with the telecom company dataset [8]. The other paper by B.Q Huang et al proposed new features set and new window technique for the customer attrition prediction in the landline telecommunications. This paper compares the proposed techniques with the existing techniques and proposed model outperformed. Finally, the use of data mining techniques can help the telecom company to improve their customer retention and profitability [9].

Mehpara saghir et al research paper with ensemble neural network model to find customer turnover in the telecom industry. The individual models include feedforward neural network technique, recurrent neural network method and radial basis function networks which includes bagging, boosting and stacking. Finally, this paper concludes the ensemble models outperformed the individual models in terms of accuracy [10]. The paper by Ning Lu et al proposes a churn forecasting model using boosting algorithms such as Adaboost, Gradient and XGBoost and compared the performance of these models with the telecom dataset and found that the performance of XGBoost was better. These papers explain about each method with real time dataset and how it helps the telecom companies to keep the existing customers and enhance their services [11].

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

The Ali Tamaddoni Jahromi et al, the authors proposed a model to predict the churn in telecom industry in non-contractual setting. Using the predictive model, it is found that the Neural network performance was good [12]. Dimensionality reduction paper by Maha Alkhayrat et al proposed a comparative study of deep learning and Principal Component Analysis (PCA) for the telecom customer segmentation and it outperformed and provided a more interpretable representation of the customer segments [13]. Another survey conducted by Preeti K delve et al (2018) examined various algorithms and customer churn in various industries which discussed the strengths and weaknesses of different Machine learning algorithms and concluded through the survey that the accuracy of the prediction model is highly dependent on the nature of the data with the specific algorithms [14].

A study by Kiran Dahiya et al (2018) study compared several machine learning algorithms for the churn prediction and the random forest performance was good [15] among the others. The recent study of Irfan Ullah et al (2022) used RNN (Recurrent Neural Networks) with a novel feature selection strategy with the Pakistani telecom company dataset where GBM (Gradient Boosting Machine) and Random Forest was implemented and evaluated [16].

The research paper by Almugren et al, explores the use of twitter data mining to anticipate customer turnover behavior in the Arabic telecom market. This paper “An empirical study on customer churn behavior using twitter mining approach” has been employed to address the issues using social media [17].

Every study shows the various algorithms with different dataset of churn customer prediction were implemented and it is clearly understandable that the feature selection is a crucial step in building a right model.

EXPERIMENT

In this Research, eight methodologies were used totally to forecast the customer attrition in Telecom industry. CatBoosting Classifier, Gradient Boosting, XGBClassifier, AdaBoost Classifier, Decision Tree, Random Forest, Logistic Regression and K-Neighbors Classifier were used to create a model using the original dataset (imbalanced dataset) and the balanced dataset was evaluated using python programming tool to find the customer attrition problems in telecom Industry.

THE DATASET

The IBM Accelerator Catalog dataset "Telco Customer Churn" which includes details about customers who have left the telecom company every month. This dataset consists of 7047 records and thirty-three columns which includes target variable "Churn Value” which displays if the customer has quit the industry or not.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

	CustomerID	Count	Country	State	City	Zip Code	Lat Long	Latitude	Longitude	Gender	...	Contract	Paperless Billing	Payment Method	Monthly Charges	Total Charges
0	3668-QPYBK	1	United States	California	Los Angeles	90003	33.964131, -118.272783	33.964131	-118.272783	Male	...	Month-to-month	Yes	Mailed check	53.85	108.15
1	9237-HQITU	1	United States	California	Los Angeles	90005	34.059281, -118.30742	34.059281	-118.307420	Female	...	Month-to-month	Yes	Electronic check	70.70	151.65
2	9305-CDSKC	1	United States	California	Los Angeles	90006	34.048013, -118.293953	34.048013	-118.293953	Female	...	Month-to-month	Yes	Electronic check	99.65	820.5
3	7892-POOKP	1	United States	California	Los Angeles	90010	34.062125, -118.315709	34.062125	-118.315709	Female	...	Month-to-month	Yes	Electronic check	104.80	3046.05
4	0280-XJGEX	1	United States	California	Los Angeles	90015	34.039224, -118.266293	34.039224	-118.266293	Male	...	Month-to-month	Yes	Bank transfer (automatic)	103.70	5036.3

5 rows × 33 columns

	City	Zip Code	Lat Long	Latitude	Longitude	Gender	...	Contract	Paperless Billing	Payment Method	Monthly Charges	Total Charges	Churn Label	Churn Value	Churn Score	CLTV	Churn Reason
	Los Angeles	90003	33.964131, -118.272783	33.964131	-118.272783	Male	...	Month-to-month	Yes	Mailed check	53.85	108.15	Yes	1	86	3239	Competitor made better offer
	Los Angeles	90005	34.059281, -118.30742	34.059281	-118.307420	Female	...	Month-to-month	Yes	Electronic check	70.70	151.65	Yes	1	67	2701	Moved
	Los Angeles	90006	34.048013, -118.293953	34.048013	-118.293953	Female	...	Month-to-month	Yes	Electronic check	99.65	820.5	Yes	1	86	5372	Moved
	Los Angeles	90010	34.062125, -118.315709	34.062125	-118.315709	Female	...	Month-to-month	Yes	Electronic check	104.80	3046.05	Yes	1	84	5003	Moved
	Los Angeles	90015	34.039224, -118.266293	34.039224	-118.266293	Male	...	Month-to-month	Yes	Bank transfer (automatic)	103.70	5036.3	Yes	1	89	5340	Competitor had better devices

Fig: The dataset

THE DATASET COLUMNS DESCRIPTION

The Dataset column sections are mentioned below. The source of Dataset column details screenshots from the IBM sample fictional Dataset [18].

Demographics

CustomerID: A unique ID that identifies each customer.

Count: A value used in reporting/dashboarding to sum up the number of customers in a filtered set.

Gender: The customer's gender: Male, Female

Age: The customer's current age, in years, at the time the fiscal quarter ended.

Senior Citizen: Indicates if the customer is 65 or older: Yes, No

Married: Indicates if the customer is married: Yes, No

Dependents: Indicates if the customer lives with any dependents: Yes, No. Dependents could be children, parents, grandparents, etc.

Number of Dependents: Indicates the number of dependents that live with the customer.

Fig: Customer Demographics details in the dataset

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

Location

CustomerID: A unique ID that identifies each customer.

Count: A value used in reporting/dashboarding to sum up the number of customers in a filtered set.

Country: The country of the customer's primary residence.

State: The state of the customer's primary residence.

City: The city of the customer's primary residence.

Zip Code: The zip code of the customer's primary residence.

Lat Long: The combined latitude and longitude of the customer's primary residence.

Latitude: The latitude of the customer's primary residence.

Longitude: The longitude of the customer's primary residence.

Fig: Customer Location details in the dataset

Services

CustomerID: A unique ID that identifies each customer.

Count: A value used in reporting/dashboarding to sum up the number of customers in a filtered set.

Quarter: The fiscal quarter that the data has been derived from (e.g. Q3).

Referred a Friend: Indicates if the customer has ever referred a friend or family member to this company: Yes, No

Number of Referrals: Indicates the number of referrals to date that the customer has made.

Tenure in Months: Indicates the total amount of months that the customer has been with the company by the end of the quarter specified above.

Offer: Identifies the last marketing offer that the customer accepted, if applicable. Values include None, Offer A, Offer B, Offer C, Offer D, and Offer E.

Phone Service: Indicates if the customer subscribes to home phone service with the company: Yes, No

Avg Monthly Long Distance Charges: Indicates the customer's average long distance charges, calculated to the end of the quarter specified above.

Multiple Lines: Indicates if the customer subscribes to multiple telephone lines with the company: Yes, No

Internet Service: Indicates if the customer subscribes to Internet service with the company: No, DSL, Fiber Optic, Cable.

Avg Monthly GB Download: Indicates the customer's average download volume in gigabytes, calculated to the end of the quarter specified above.

Online Security: Indicates if the customer subscribes to an additional online security service provided by the company: Yes, No

Online Backup: Indicates if the customer subscribes to an additional online backup service provided by the company: Yes, No

Device Protection Plan: Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company: Yes, No

Fig: Service details in the dataset

Premium Tech Support: Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times: Yes, No

Streaming TV: Indicates if the customer uses their Internet service to stream television programing from a third party provider: Yes, No. The company does not charge an additional fee for this service.

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.

Streaming Music: Indicates if the customer uses their Internet service to stream music from a third party provider: Yes, No. The company does not charge an additional fee for this service.

Unlimited Data: Indicates if the customer has paid an additional monthly fee to have unlimited data downloads/uploads: Yes, No

Contract: Indicates the customer's current contract type: Month-to-Month, One Year, Two Year.

Paperless Billing: Indicates if the customer has chosen paperless billing: Yes, No

Payment Method: Indicates how the customer pays their bill: Bank Withdrawal, Credit Card, Mailed Check

Monthly Charge: Indicates the customer's current total monthly charge for all their services from the company.

Total Charges: Indicates the customer's total charges, calculated to the end of the quarter specified above.

Fig: Service Details in the dataset

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

Churn Label: Yes = the customer left the company this quarter. No = the customer remained with the company. Directly related to Churn Value.

Churn Value: 1 = the customer left the company this quarter. 0 = the customer remained with the company. Directly related to Churn Label.

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.

Churn Score Category: A calculation that assigns a Churn Score to one of the following categories: 0-10, 11-20, 21-30, 31-40, 41-50, 51-60, 61-70, 71-80, 81-90, and 91-100

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.

CLTV Category: A calculation that assigns a CLTV value to one of the following categories: 2000-2500, 2501-3000, 3001-3500, 3501-4000, 4001-4500, 4501-5000, 5001-5500, 5501-6000, 6001-6500, and 6501-7000.

Churn Category: A high-level category for the customer's reason for churning: Attitude, Competitor, Dissatisfaction, Other, Price. When they leave the company, all customers are asked about their reasons for leaving. Directly related to Churn Reason.

Churn Reason: A customer's specific reason for leaving the company. Directly related to Churn Category.

Fig: Churn Details in the dataset

The selected dataset has both categorical and numerical data. It is widely used in machine learning projects to predict customer attrition in the telecom industry to regenerate response.

EXPLORING THE DATASET

This dataset has 7043 rows with thirty-three columns and "Churn Label / Churn Value" is the target feature in the dataset along with the display of two distinct values "Yes" and "No." This gives an idea of the percentage of customers who have churned and percentage of customers who have not. From the Dataset, to conclude that out of 7043 rows in the dataset, 1869 customers have churned and 5174 customers are the existing.

```
In [4]: #analysing Target feature Count
        dataset["Churn Label"].value_counts()

Out[4]: No     5174
        Yes    1869
        Name: Churn Label, dtype: int64
```

Fig: Target Feature Count

Duplicate rows have the potential to result in inaccurate outputs. When the dataset was verified for duplicates, there were no duplicate records in the dataset.

```
In [6]: #Check for Duplicates
        dataset.duplicated().sum()

Out[6]: 0
```

Fig: Checking Duplicates in the dataset

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

The statistics summary in the numerical columns provide a wide range of information related to the range and distribution of every single numerical column. For every single column, the statistics provide mean, count, standard deviation and lowest & highest values.

```
In [7]: #Display summary statistics for a dataset
        dataset.describe()
```

Out[7]:

	Count	Zip Code	Latitude	Longitude	Tenure Months	Monthly Charges	Churn Value	Churn Score	CLTV
count	7043.0	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000
mean	1.0	93521.964646	36.282441	-119.798880	32.371149	64.761692	0.265370	58.699418	4400.295755
std	0.0	1865.794555	2.455723	2.157889	24.559481	30.090047	0.441561	21.525131	1183.057152
min	1.0	90001.000000	32.555828	-124.301372	0.000000	18.250000	0.000000	5.000000	2003.000000
25%	1.0	92102.000000	34.030915	-121.815412	9.000000	35.500000	0.000000	40.000000	3469.000000
50%	1.0	93552.000000	36.391777	-119.730885	29.000000	70.350000	0.000000	61.000000	4527.000000
75%	1.0	95351.000000	38.224869	-118.043237	55.000000	89.850000	1.000000	75.000000	5380.500000
max	1.0	96161.000000	41.962127	-114.192901	72.000000	118.750000	1.000000	100.000000	6500.000000

Fig: Statistics summary of the dataset

As per the summary statistics, the gathered information is that a customer exists in a company with an average tenure of 32.37 months, with a standard deviation of 24.55 months. The monthly fee paid by customers is an average of 64.76. The churn score, which examines how a client / customer is about to stop using the services offered by company runs from 5 to 100 which states that 26% of the customers have stopped using company's services. These clients have an average churn rate of 58.69 percent.

Null values, which are also known as missing data which has a greater impact which leads to inaccurate outputs. As per the research, "Churn Reason" has the highest number of null values (73.46%). This information will help to manage the missing data and ways to rectify them.

```
In [9]: (dataset.isnull().mean().sort_values(ascending=False)[0:33])*100
Out[9]: Churn Reason       73.463013
        Online Security     0.000000
```

Fig: Checking Null values in the dataset

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

DATA CLEANING

Data cleaning is the process which prepares datasets for statistical analysis. To assure that the dataset is fine. Data cleaning such as dropping the extra columns, managing the missing data and checking for duplicates in the dataset.

The 'Count' column is exceptional for analysis. Hence, the ‘Count' column is removed from the dataset. The 'Total Charges' column does not have numerical values in all the rows and the operator (~) used in the coding to check the non-numerical value in it.

The 'Total Charges' column has a few non numerical values which have been replaced using the mean value of the column. This is mandatory because the missing info will lead to incorrect results and replacing them with mean is the best way to deal with the missing information.

Out[13]:

City	Zip Code	Lat Long	Latitude	Longitude	Gender	Senior Citizen	...	Contract	Paperless Billing	Payment Method	Monthly Charges	Total Charges	Churn Label	Churn Value	Churn Score	CLTV	Churn Reason
San rdino	92408	34.084909, -117.258107	34.084909	-117.258107	Female	No	...	Two year	Yes	Bank transfer (automatic)	52.55		No	0	36	2578	NaN
lence	93526	36.869584, -118.189241	36.869584	-118.189241	Male	No	...	Two year	No	Mailed check	20.25		No	0	68	5504	NaN
Mateo	94401	37.590421, -122.306467	37.590421	-122.306467	Female	No	...	Two year	No	Mailed check	80.85		No	0	45	2048	NaN
ertino	95014	37.306612, -122.080621	37.306612	-122.080621	Male	No	...	Two year	No	Mailed check	25.75		No	0	48	4950	NaN
lcrest	95569	40.363446, -123.835041	40.363446	-123.835041	Female	No	...	Two year	No	Credit card (automatic)	56.05		No	0	30	4740	NaN
geles	90029	34.089953, -118.294824	34.089953	-118.294824	Male	No	...	Two year	No	Mailed check	19.85		No	0	53	2019	NaN
n City	92585	33.739412, -117.173334	33.739412	-117.173334	Male	No	...	Two year	No	Mailed check	25.35		No	0	49	2299	NaN
mond	95005	37.078873, -122.090386	37.078873	-122.090386	Female	No	...	Two year	No	Mailed check	20.00		No	0	27	3763	NaN
Verne	91750	34.144703, -117.770299	34.144703	-117.770299	Male	No	...	One year	Yes	Mailed check	19.70		No	0	69	4890	NaN
Bell	90201	33.970343, -118.171368	33.970343	-118.171368	Female	No	...	Two year	No	Mailed check	73.35		No	0	44	2342	NaN
ngton	90744	33.782068, -118.262263	33.782068	-118.262263	Male	No	...	Two year	Yes	Bank transfer (automatic)	61.90		No	0	65	5188	NaN

Fig: Non numerical values Total charges column

The null value in each column is then printed which will be used for finding and dealing with missing data. As 'Churn Reason' column has 5174 null values saying that this information is lost for a major part of clients.

Eventually the data frame is checked to list if there are any duplicate rows. Duplicate data will lead to misinformation and it is mandatory to check at the first stage. In this scenario, there are no such duplicate rows, so the next step was conducted forward for the process of analysis.

```
In [15]: # Checking for duplicates
         dataset.duplicated().sum()

Out[15]: 0
```

Fig: Checking duplicates in the dataset

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

EXPLORATORY DATA ANALYSIS

To examine this dataset, the columns of IBM Telecom customer churn dataset were divided into two different categories: "numeric_features" and "categorical_features".

```
We have 9 numerical features : ['Zip Code', 'Latitude', 'Longitude', 'Tenure Months', 'Monthly Charges', 'Total Charges', 'Chur
n Value', 'Churn Score', 'CLTV']

We have 23 categorical features : ['CustomerID', 'Country', 'State', 'City', 'Lat Long', 'Gender', 'Senior Citizen', 'Partner',
'Dependents', 'Phone Service', 'Multiple Lines', 'Internet Service', 'Online Security', 'Online Backup', 'Device Protection',
'Tech Support', 'Streaming TV', 'Streaming Movies', 'Contract', 'Paperless Billing', 'Payment Method', 'Churn Label', 'Churn Re
ason']
```

Fig: List of Categorical and Numerical Features

In the above figure there are thirty-two columns in the dataset, twenty-three of them are categorized and nine are numerical.

There are some lists of features included in numerical such as “Zip Code”, "Latitude", "Longitude", "Tenure Months", "Monthly Charges", "Total Charges", "Churn Value", "Churn Score", and "CLTV". These characteristics have a calculable nature and are calculated by numbers. For Example, "Tenure Months" describes the number of months the customers existed in the business and the "Monthly Charges" describes the amount that has been charged monthly.

The Clear-cut feature on the other hand includes “Customer ID”, “Country”, “State”, “Lat Long”, “Gender”, “Citizenship”, “partner”, “Dependents”, “Phone services”, “Multiple Lines”, “Internet service”, “Online security” , “Backup”, “streaming TV with Movies” and “paperless billing”. These features are not calculative in nature and they do not have any numerical value to describe the same. For Example, the feature ‘Gender’ is taken into consideration which has two types such as male and female. For this use, the term “categorical feature “is taken as a non- numeric value to describe the same.

During the process of Cleaning the data, preprocessing and Modelling, it is important to understand the various aspects of dataset including each feature in detail. In the case of Numerical features, the scaling method may be required and Label encoding/ one hot encoding may be necessary for the Categorical features. The machine learning method can be selected based on the problem statement and situation as it is showing the ‘risk clients’ with retention strategy.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

CHECKING THE PROPORTIONS OF CATEGORICAL DATA

The findings from the proportion of the categorical data show that there is only one value for the State and Country column for all the customers. The internet related services like Online Backup service, Device Protection, Online security, Tech help, Streaming TV and Movies are not activated by many customers in the dataset. A minor majority of customers are male and most of the male customer has “Month-to-Month” Contract. Almost 26.5 % of customers have left the company and most of the customers are using “Electronic Check” payment method.

```
Name: Senior Citizen, dtype: float64
----------------------------
No     51.69672
Yes    48.30328
Name: Partner, dtype: float64
----------------------------
No     76.899049
Yes    23.100951
Name: Dependents, dtype: float64
----------------------------
Yes    90.316626
No      9.683374
Name: Phone Service, dtype: float64
----------------------------
No                  48.132898
Yes                 42.183729
No phone service     9.683374
Name: Multiple Lines, dtype: float64
```

```
Name: Online Backup, dtype: float64
----------------------------
No                     43.944342
Yes                    34.388755
No internet service    21.666903
Name: Device Protection, dtype: float64
----------------------------
No                     49.311373
Yes                    29.021724
No internet service    21.666903
Name: Tech Support, dtype: float64
----------------------------
No                     39.897771
Yes                    38.435326
No internet service    21.666903
Name: Streaming TV, dtype: float64
----------------------------
No                     39.542808
Yes                    38.790288
```

```
Name: Streaming Movies, dtype: float64
----------------------------
Month-to-month    55.019168
Two year          24.066449
One year          20.914383
Name: Contract, dtype: float64
----------------------------
Yes    59.221922
No     40.778078
Name: Paperless Billing, dtype: float64
----------------------------
Electronic check             33.579441
Mailed check                 22.887974
Bank transfer (automatic)    21.922476
Credit card (automatic)      21.610109
Name: Payment Method, dtype: float64
----------------------------
No     73.463013
Yes    26.536987
```

Fig: proportions of categorical data

The “CustomerID”, “City” and “Lat Long” Columns are considered as high in cardinality which makes it difficult to derive the important insights. So, it is better not to consider these columns in further processes.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

UNIVARIATE ANALYSIS

NUMERICAL FEATURES

The below Histogram of the Telco dataset describes the numerical attribute distribution. Histograms used have been shown in the ensuing figure. The “X- axis” describes the values of the data and the “Y- axis” describes the frequency of those values. Distribution of values for a particular feature has been described in each histogram.

A small set of customers have greater charges when compared to other major groups of customers. This can be seen in the “Total Charges” graph below. The distribution of monthly charges seems very normal raising at the range of 70-80. But the distributions of churn score, churn value and CLTV are dynamic which shows the presence of few customers with the high values.

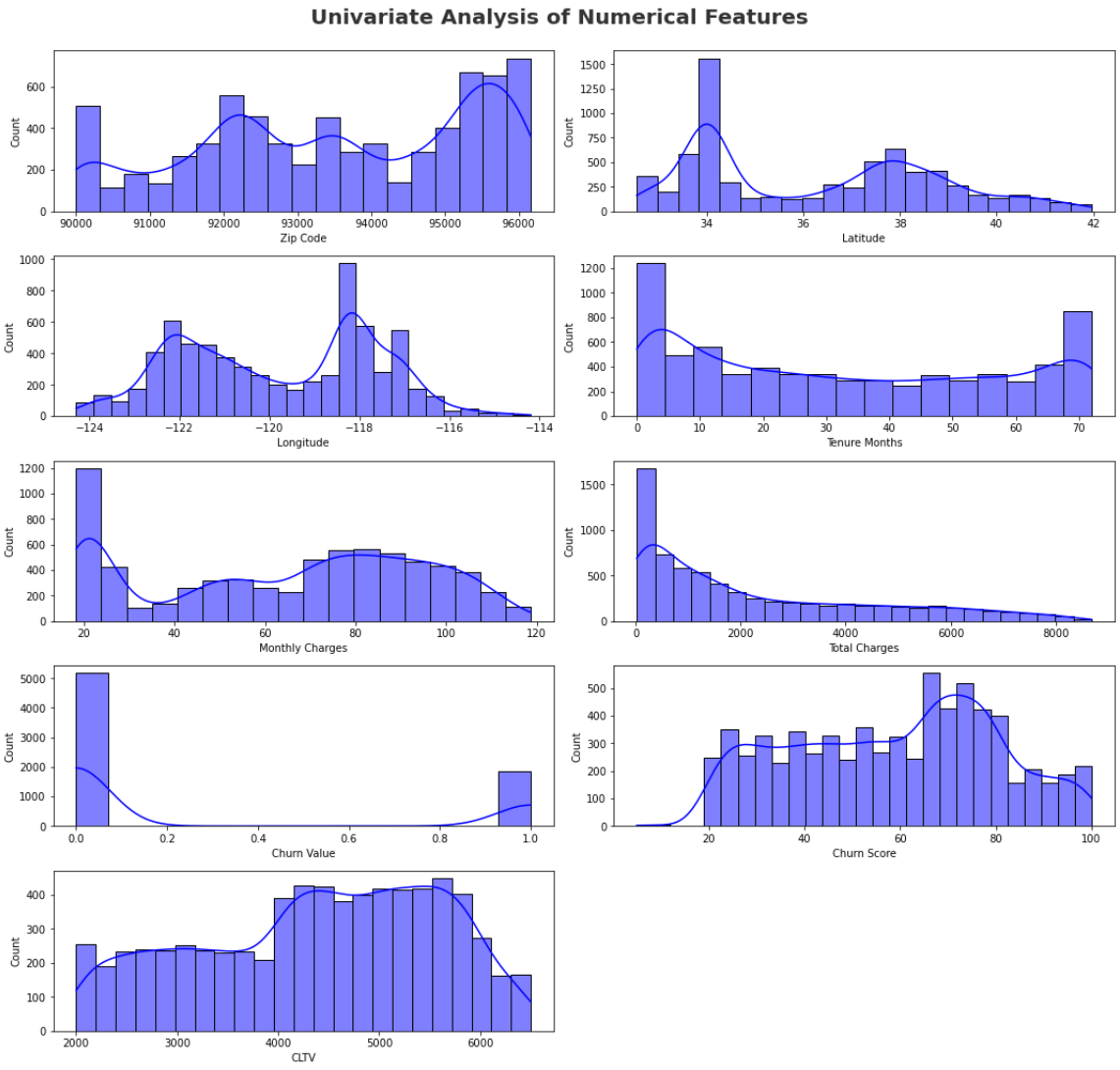


Fig: Univariate Analysis of Numeric Features

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

CATEGORICAL FEATURES

Categorical univariate analysis is beneficial in finding the missing values which can be taken further and cleaned which can be taken into consideration if necessary. Also, it helps in showing the non-relevant which can be removed from the dataset which in turn makes the analysis process quite quicker and easier.

Country: The below figure describes the percentage of clients in each nation/country. United States citizens ranks first in clientele followed by British and Indian citizens.

State: The below figure describes the percentage of clients from each state. California holds the highest percentage of customers followed by Texas and New York.

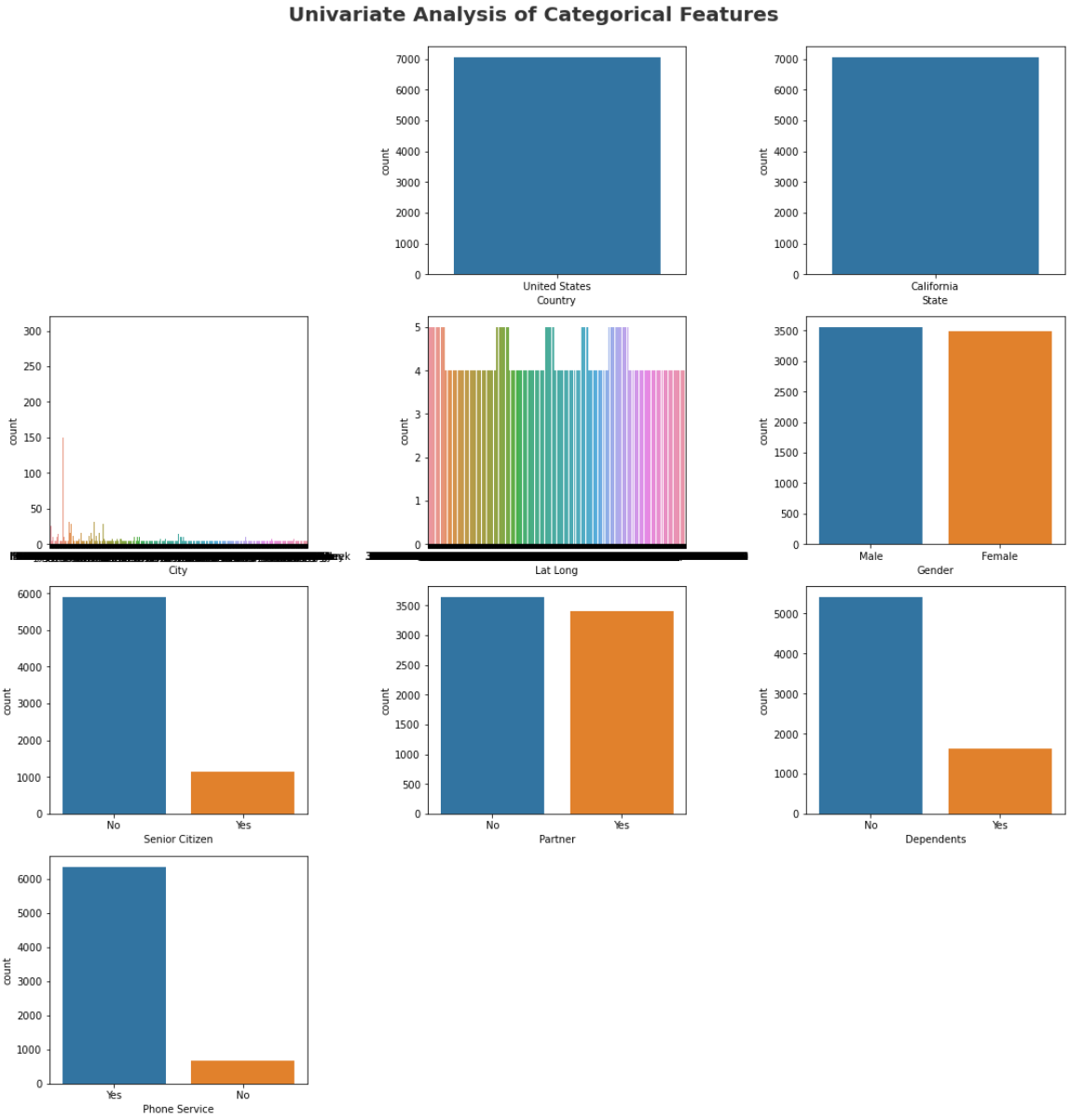


Fig: Univariate Analysis for Categorical Features

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

Gender: The percentage of the male and female is portrayed in the above graph. They are equal in the dataset.

Partner: The percentage of the customers who have a partner has been mentioned in the graph. It is the same for customers who have partners and those who do not have partners.

Dependents: The percentage of the customers who have dependents are mentioned in the plot which shows that many customers do not have dependents in the dataset.

Phone Service: The part of the clients who have phone services has been mentioned in the graph. Most of the customers have phone services as per the dataset.

MULTIVARIATE ANALYSIS

Data that is not noticeably clear while analyzing the variables can create complicated problems and relationships which can only be cleared using multivariate analysis. Multivariate analysis helps in the development of forecasting models, making data driven decisions and understanding the underlying structure of data.

```
We have 0 discrete features : []

We have 8 continuous_features : ['Zip Code', 'Latitude', 'Longitude', 'Tenure Months', 'Monthly Charges', 'Total Charges', 'Chu
rn Score', 'CLTV']

We have 1 encoded_categorical : ['Churn Value']
```

Fig: Categorizing the Numeric Feature

There are no such discrete features in the dataset. Zip Code which is known as postcode, Latitude, Longitude, Total Tenure Months, Total Monthly Charges, Total Charges, Customer Churn Score and CLTV are the eight continuous features. Only these factors will accept any kind of value which comes within their ranges and holds a vast range of numerical values.

'Churn Value' is the only present categorical feature that has been encoded. This factor has two different values which indicate whether the customer is churned or not: 0 and 1. It has been encoded as a numerical number to make the machine learning algorithm process easier.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

CHECK MULTICOLLINEARITY IN NUMERICAL FEATURES

The Correlation coefficients between numerical features has been described in the below graph. Numerical values which have one state positive correlation while values which hold -1 state negative correlation. Values that hold Zero means that there is no such correlation.

	Latitude	Longitude	Tenure Months	Monthly Charges	Total Charges	Churn Score	CLTV
Latitude	1.000000	-0.876779	-0.001631	-0.019899	-0.010307	-0.007684	0.000886
Longitude	-0.876779	1.000000	-0.001678	0.024098	0.009039	0.004260	0.000485
Tenure Months	-0.001631	-0.001678	1.000000	0.247900	0.824757	-0.224987	0.396406
Monthly Charges	-0.019899	0.024098	0.247900	1.000000	0.650468	0.133754	0.098693
Total Charges	-0.010307	0.009039	0.824757	0.650468	1.000000	-0.124251	0.341384
Churn Score	-0.007684	0.004260	-0.224987	0.133754	-0.124251	1.000000	-0.079782
CLTV	0.000886	0.000485	0.396406	0.098693	0.341384	-0.079782	1.000000

Fig: Correlation Matrix for Numeric Features

The Latitude and Longitude have a negative correlation (-0.877) when seen in the table which states geographic coordinates which should be related inversely. Most of the values fall between -0.2 and 0.2 and the remaining features are categorized as weak. The largest possible value (0.825) falls between Tenure Months and Total Charges which means that a longer Tenure would result in higher Total Charges.

CORRELATION MATRIX

As discussed, 'Total Charges' and 'Tenure Months' hold the highest positive correlation which states that the charges would increase based on the duration of the customer’s subscription and accordingly the heatmap.

As expected, the correlation between "Monthly Charges" and "Total Charges" turned out to be positive. When given the inverse relationship between "Churn Score" and "Tenure Months", shows that customers with higher churn rates stay only for a shorter period in the business.

Additionally, "Churn Score" and "Total Charges" has been correlated negatively states that clients with higher churn rate will earn a less amount of profit. Overall, this heatmap helps to identify the features that relate to one another and how they are related to each other.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

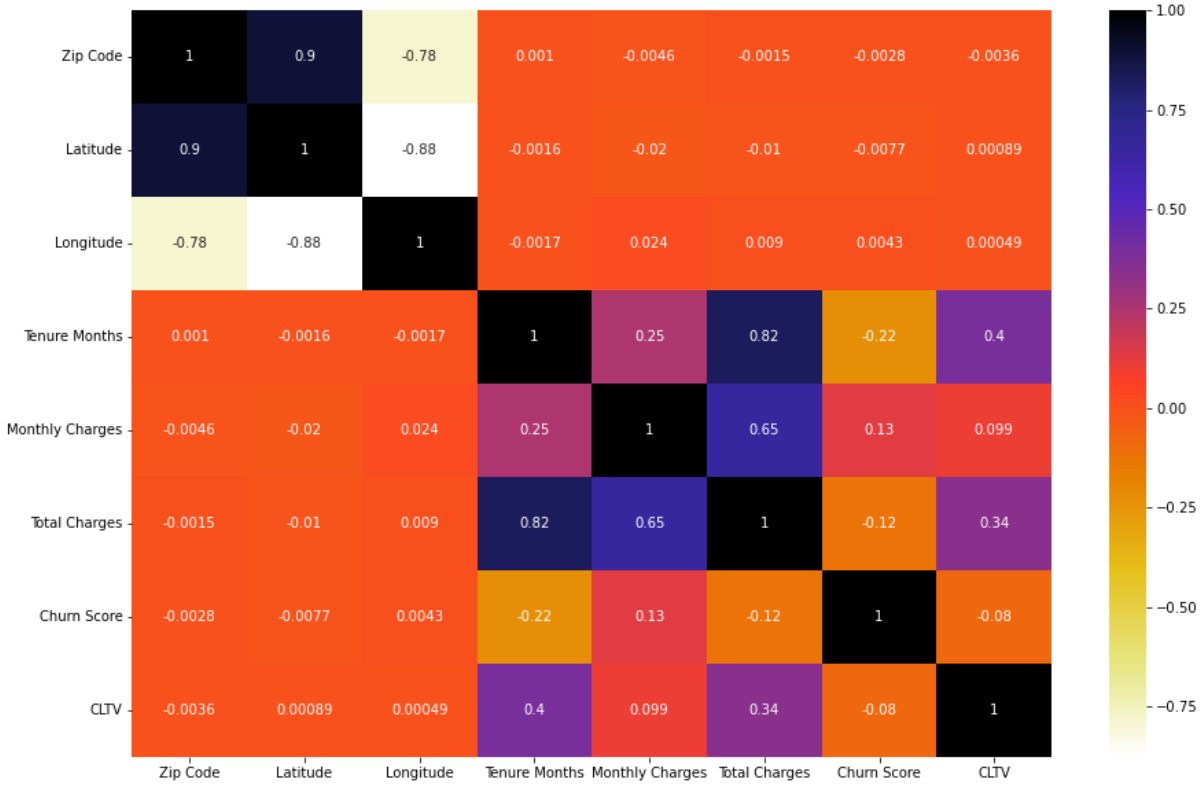


Fig: Heatmap to Understand Correlation

CHECK MULTICOLLINEARITY IN CATEGORICAL FEATURES

To check whether the target feature "Churn Value" and the dataset categories has been connected, it becomes mandatory to conduct a chi-squared test of independence. The Chi-squared test does not consider the null hypothesis and the other two properties which are independent. If the value of p is less than 0.05, it states that there is a correlation between the two factors.

Out of twenty-three categorical factors, seventeen features have a significant correlation with “Churn Value”. As per the result, p-values are less than 0.5 which consists of city, Senior Citizen, Partner Status, Dependents, Multiple lines of services, Internet service, Online security, Backup, Device Protection which is security of the device, Paperless Billing, Payment Method, Churn Label. The other six features such as Phone service, Customer ID, Gender, State, Lat long and Country are independent and it shows that they are not connected strongly.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

OUTLIER ANALYSIS

In the Initial stage of outlier analysis, the columns like “CustomerID” are removed which does not provide any relevant information to the analysis. By using select_dtypes method, choose the column which has numerical values and assign the name for the columns(num_data).

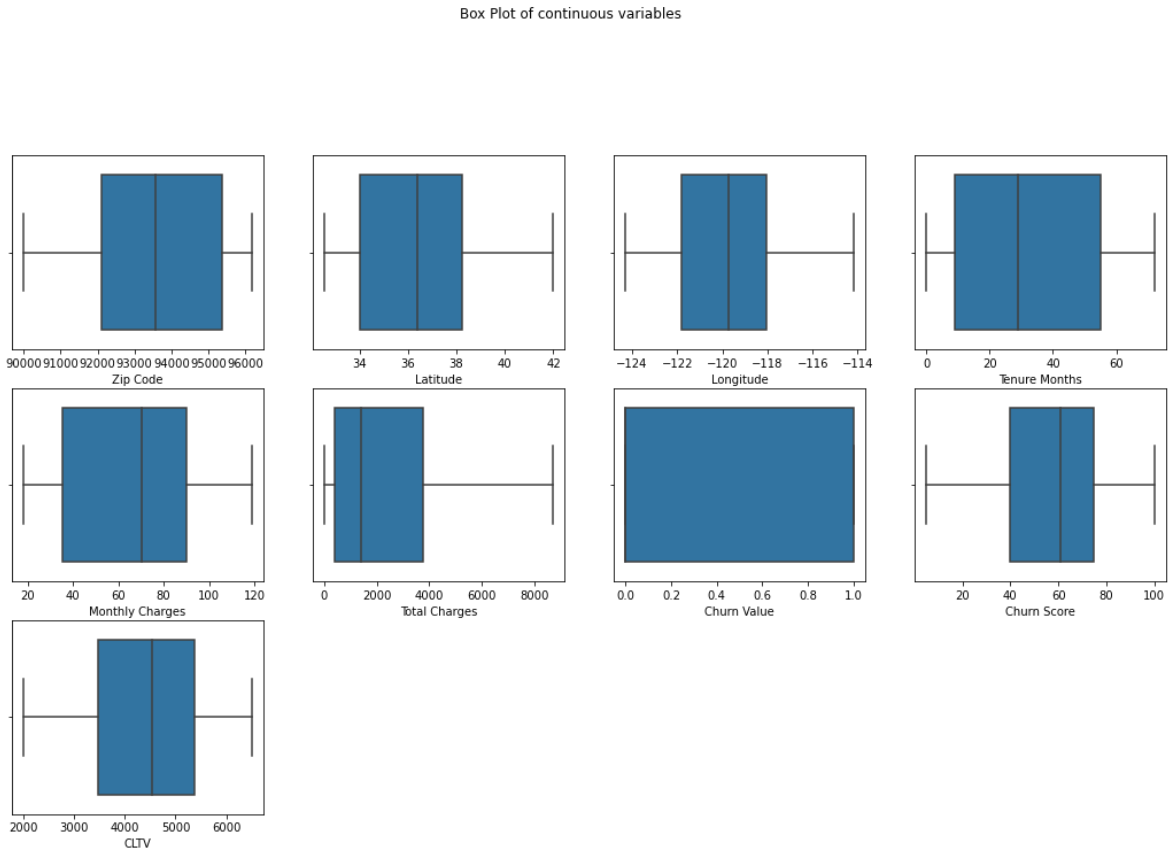


Fig: Box Plot for Outlier Analysis

Continuous variables in the dataset are displayed in the plot. The interquartile range (IQR), the range of the data and others are demonstrated as points outside the whiskers in the box plots which figure out the distribution of the data for each variable.

The variables like “Tenure Months”, “Monthly Charges” and “Total Charges” in the above box plots have no outliers. These are the data points which come under the normal distribution of various data points. They hold the ability to complicate the data which induces them to change the findings. Outliers must be managed with utmost care to avoid these kinds of circumstances. Adding to the above, major portion of customers have poor customer lifetime values (CLTV) and turnover rates. This states that customers who have lower rate of CLTV have the lowest chance to churn which has a greater impact in marketing and retention strategies.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

VISUALIZATION

EXPLORING CHURN FREQUENCY IN CUSTOMER DATA

The dispersal of customer attrition among those who hold with and without phone service has been predicted in the below graph. The customers who have phone service hold a higher rate of churn when compared to others who do not have phone service. In assumption of poor call quality, dissatisfaction with the call content and duration of the call can be considered as some of the primary features which increases the churn among the ones who hold phone services.

The possibility of the users using various phone services has been mentioned in the below pie chart. Customers who have fiber optic internet service rank high at 44%, followed by DSL at 34.4%. 21.7% of the customers do not have internet service. From the below information, services which can be provided online can be analyzed well. Marketing techniques and campaigns can also be planned accordingly.

There are various types of contracts and the percentage of the clients using those contracts has been mentioned in the pie chart. 55 % of clients hold month-month contracts, 24.1% of the clients hold two-year contracts and 20.9% of the clients hold one-year contracts. Based on the below pie chart information, discounts and promotions can be planned accordingly.

The below graph shows the clients who have various contract types. Customers who hold month-to-month contracts have a higher rate of churn when compared to customers who hold one-to-two-year contracts. Increased monthly costs and lower client satisfaction are the main factors which increase the churn rate for month-to-month contract customers.

Finally, the below-mentioned pie chart clearly mentions the customer turnover in the types of contracts and its factors affecting the same. This information can be used by businesses to take a robust decision on payment method, internet services and contracts. Based on this, every business can plan accordingly and start their techniques to lower the risk of customer churn.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

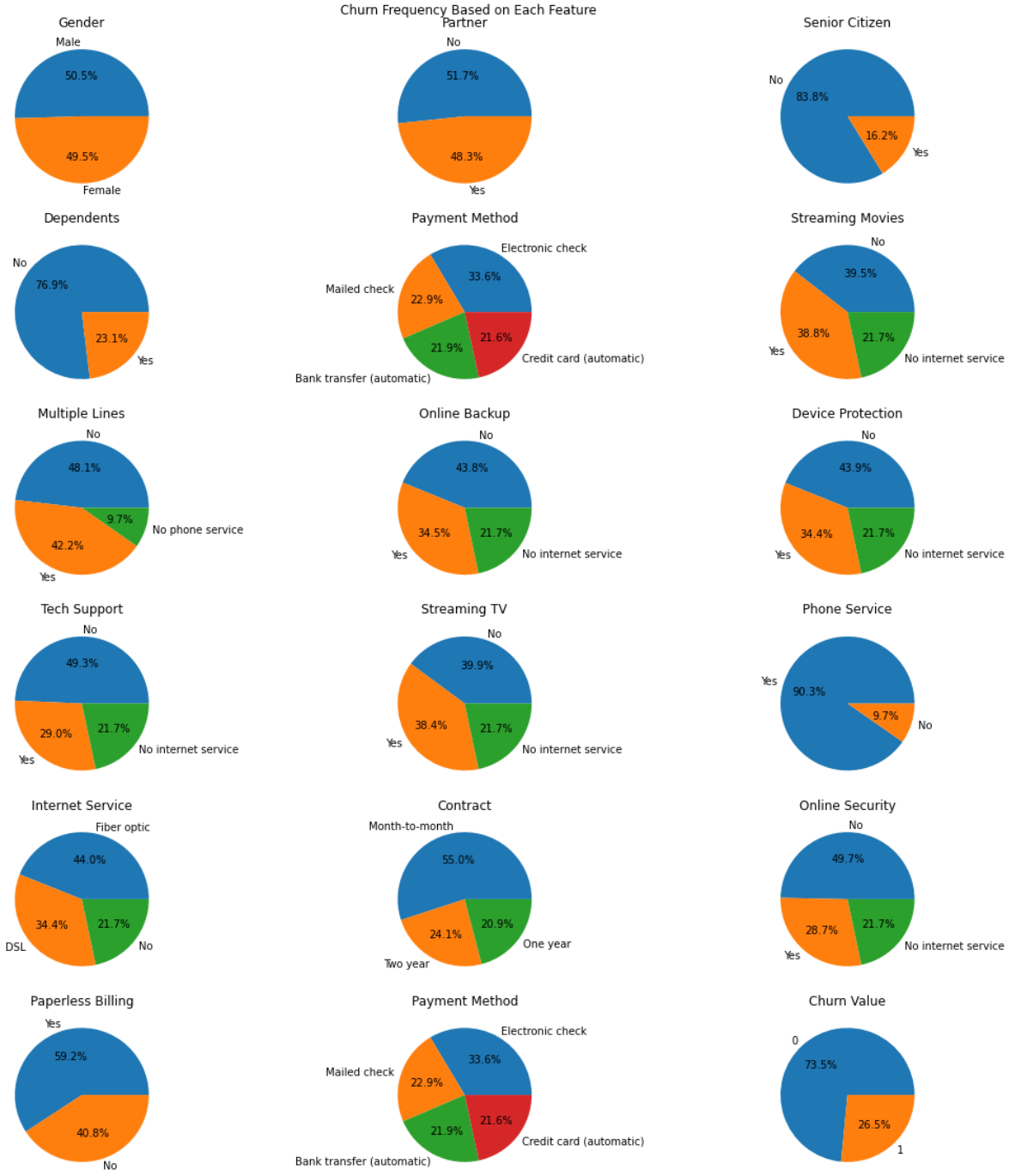


Fig: Churn Frequency based on Each Feature

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

CHURN STATUS FOR EACH OF THE CUSTOMER DEMOGRAPHICS

The bar chart below reveals the status of customer demographics turnover. The percentage of churners and non-churners is not mentioned which is the important thing to be noticed at the initial stage.

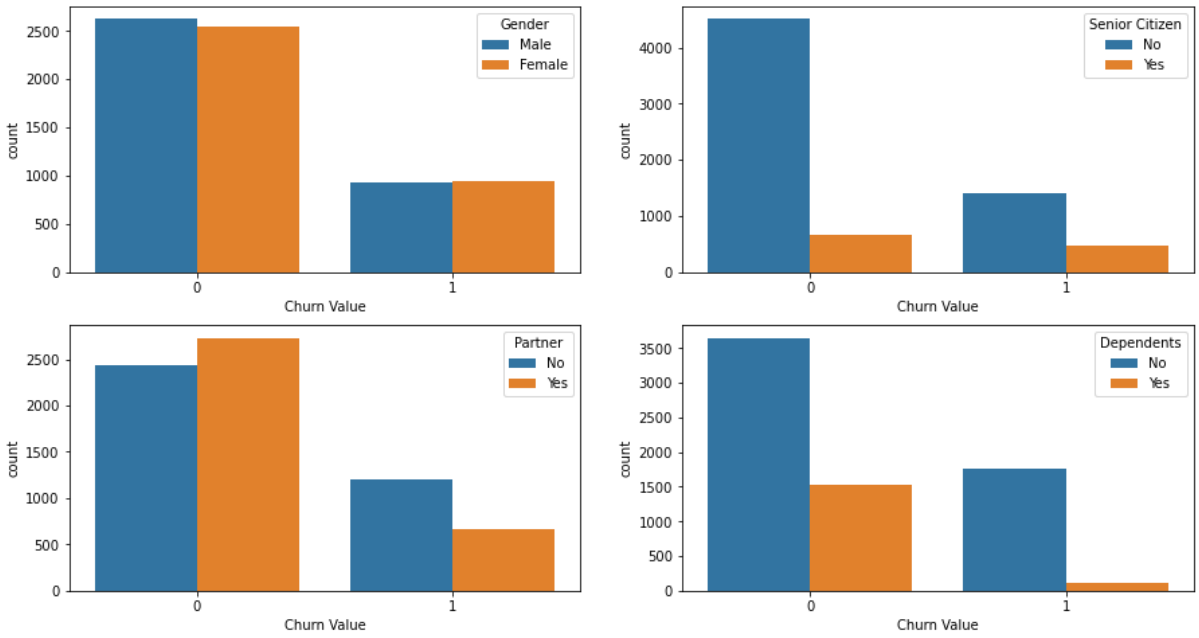


Fig: Churn based on Customer Demographics

According to the above graph, number of male and female churners has been mentioned in a predictive basis and there are slightly equal number of male churners when compared to the females. This states that the gender gap does not play a leading role in churn when compared to other major factors.

As per the Senior citizen chart, it seems like non senior citizens are most likely to churn when compared to older adults. It can be seen clearly in the chart. Age can be considered as a main factor and this proves that the age factor plays a vital role in customer attrition.

Churners exceed non-churners among customers who do not have partners. Non-churners exceed churners among consumers who have partners. So, it states that being a partner also plays a critical influence in customer attrition.

The count plot in the above-mentioned figure with a different shade states the ‘Dependents’. It clearly says that the customers without dependents are likely to have a greater rate of turnover compared to customers who have dependents. This states that children (Dependents) play a significant role in customer attrition.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

CHURN STATUS FOR EACH OF THE CUSTOMER ACCOUNT DETAILS

Based on the information, the code here has three bar charts. Each chart displays the number of churned and non- churned clients. As per the first chart, it indicates that customers who hold "Month-to-Month" contracts are more inclined to churn when compared to other customers who have long term contracts.

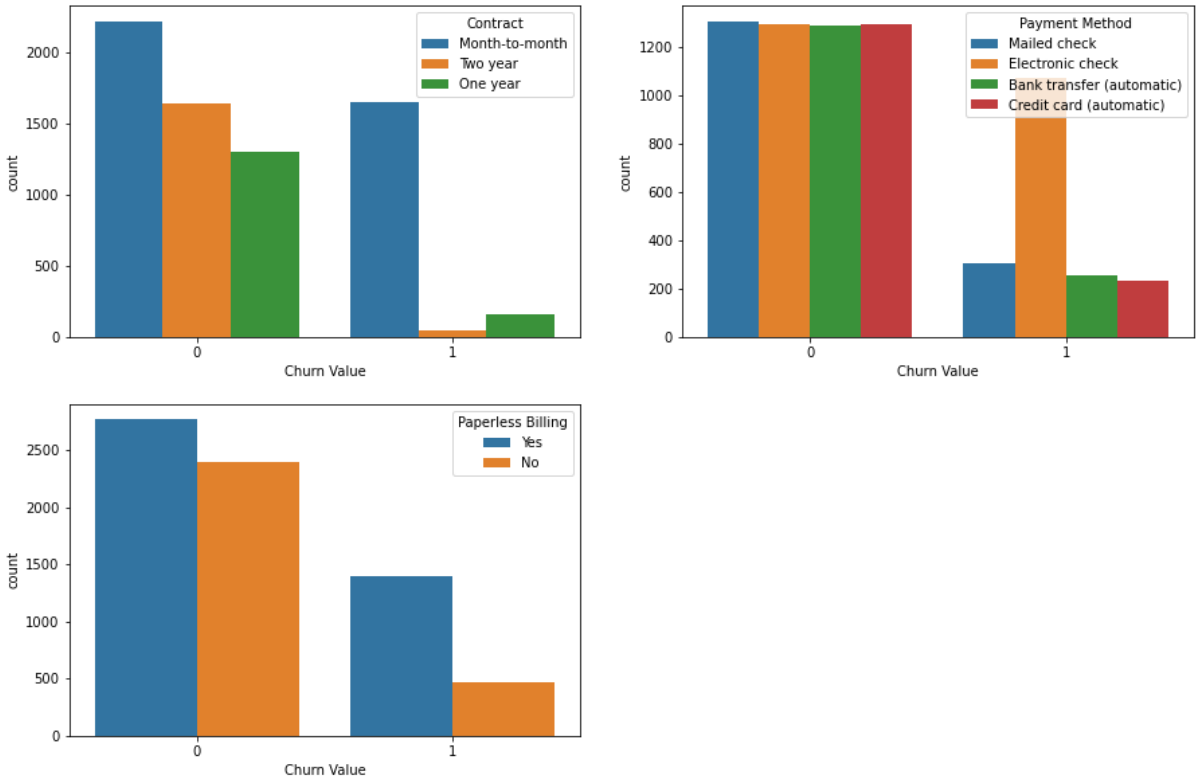


Fig: Churn based on Customer Account Details

The second chart states that the customers who pay through the mode of "electronic check" are more likely to churn when compared to other customers who prefer other payment methods.

The Third chart shows that the customers who use "paperless billing" are more probable to churn when compared to other customers who do not use the same.

As per the findings, customers who hold "month-to-month" contracts are much more likely to leave.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

CHURN STATUS FOR EACH OF THE CUSTOMER SERVICE USAGE

Some fascinating details have been uncovered by the customer utilization in the count plot below. It revealed that the churning rate of customers who hold “phone services” are high and regarding the churning rate of customers who has “multiple lines” does not churn adequately.

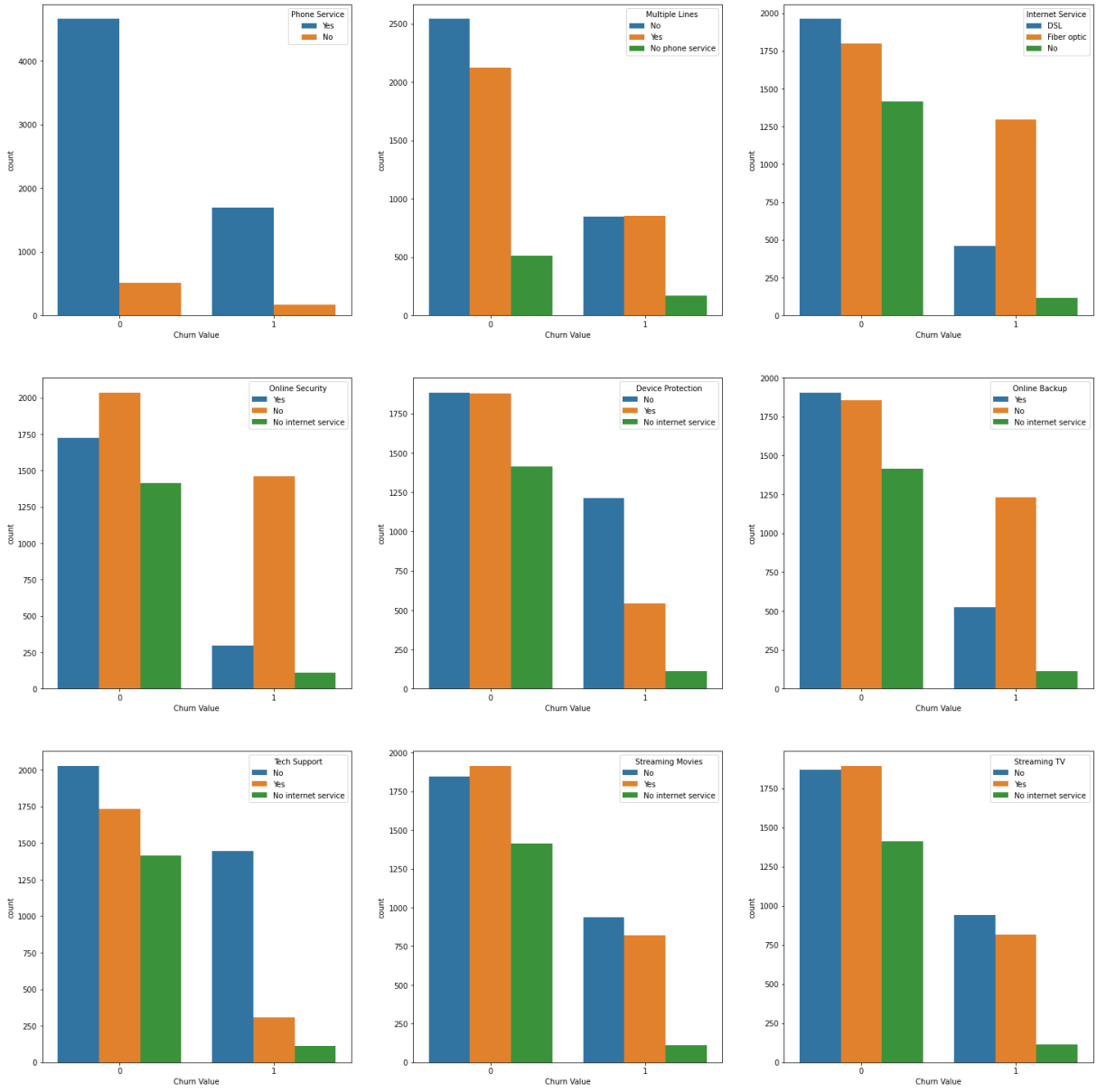


Fig: Churn based on Customer Service Usage

Customers who have fiber optic internet connection and no online security are churned high and the customers who have not activated the device protection, online backup and tech support churned more. The attrition rate of customers is equal in streaming TV and movies.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

VISUALIZATION BASED ON TOTAL CHARGES

TOTAL CHARGES BY PARTNER STATUS

As per the research and visualization, customers who do not have a partner always experience a higher total charge than the ones who already have a partner.

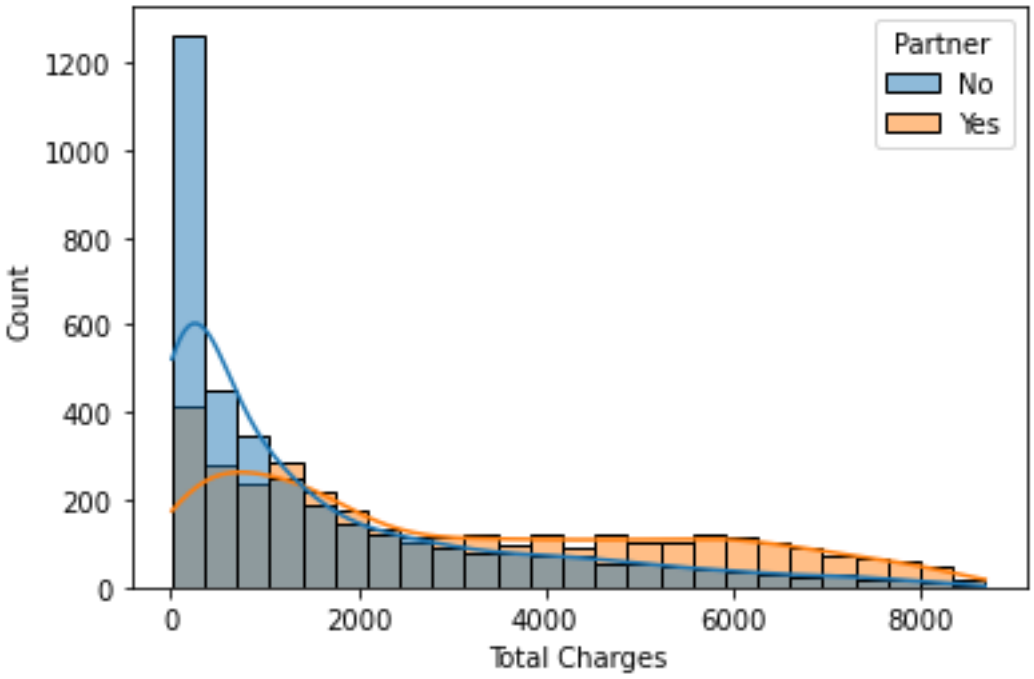


Fig: Total charges by Partner Status

Customers who have partners are charged on a higher basis when compared to the customers who do not have a partner in 0 to 5000 range. Adding to the above, customers with partners always have a higher mean when compared to others who do not have a partner. Customers without a partner always have a lower median Total Charges. This states how partners can play a vital role in forecasting the rates of customer attrition.

TOTAL CHARGES IN RELATION TO CITY

As per the below data, the total charges in Los Angeles ranks the highest at 650,034.55, followed by San Diego with 354,896.60 and Eagleville has the lowest total charge of 203.40. The below-mentioned data has been collected based on 1129 distinct cities.

The below information will help businesses to predict where they can earn more money and in which area. The telecom industry needs to concentrate on various techniques which help in attracting customers and promoting the growth of the products.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

```
City
Los Angeles       650034.550441
San Diego         354896.600000
Sacramento        256295.050000
San Jose          243735.550000
San Francisco     221624.650000
                         ...
Homeland             668.250000
Dana Point           556.800000
Loleta               484.650000
Truckee              479.350000
Eagleville           203.400000
Name: Total Charges, Length: 1129, dtype: float64
```

Fig: Total Charges by City

TOP 20 CITIES BY TOTAL CHARGES IN THE DATASET

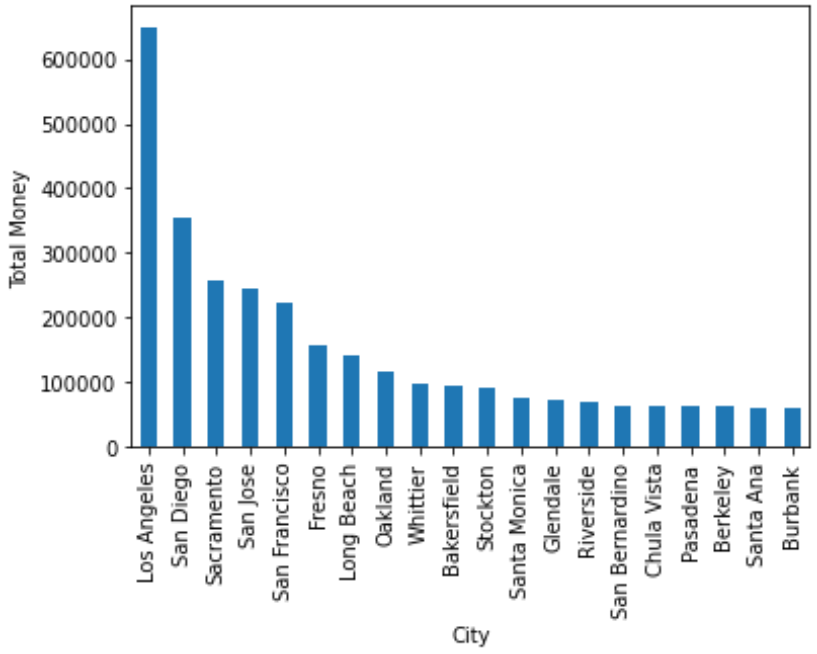


Fig: Top 20 cities by Total Charges

The above data has been merged by cities in the code says the total fees for each city. The topmost twenty cities who are charged high for the total charges have been mentioned followed by the cities who had the least total charges. From this, the telecom industry can have the knowledge of which city they can earn more income and it is about the cities which must be taken into consideration to increase productivity and decrease the customer attrition rates.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

ANALYSIS OF TOTAL CHARGES BY GENDER

As the means and medians are almost similar, this analysis clearly states that there are no such notable changes in total charges between males and females. Looking at various kinds of data, some have higher rate of total charges and some have the average rate of total charges which is shown by their standard deviation.

Gender	mean	median	std
Female	2283.191142	1389.05	2269.201601
Male	2283.407680	1406.65	2261.189740

Fig: Total Charges by Gender

DATA PRE-PROCESSING

DATA CLEANING

To ensure that the data collected is free from mistakes and incorrect information and data cleaning steps have been done when it is all set for the purpose of machine learning. The data is prepared and used for feature selection, model training and evaluation only after data cleaning has been done.

A numerous action is performed to make sure that the data collected is all set for the purpose of machine learning and started to execute the next procedure that is Total Charges to Numeric: The pd.to_numeric () methodology is used to amend the 'Total Charges' column from object data type to numerical data type.

Check Null Values: To determine the missing values in the dataset, isnull() method is used and to analyze the percentage of the missing values a for loop has been used. From this, the features were discovered where the “Total Charges” and “Churn Reason” have missing values of 0.16% and 73.46%.

```
Total Charges 0.15618 % missing values
Churn Reason 73.46301 % missing values
```

Fig: Missing Values in the Features

Handling Duplicates: Duplicates in the dataset can be analyzed through duplicated() method. But none of the duplicates in the dataset was found.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

Removing features that were not necessary for Model Training: As per the report of EDA, some features were removed which seemed to be irrelevant in the dataset. Features like “CustomerID”, “Zipcode”, “Count”, “Country”, “State”, “Lat Long”, “Latitude”, “Churn Label” and “Phone services” have been removed from the dataset using drop () technique.

FEATURE ENGINEERING

Feature Engineering is the process of choosing and extracting the relevant features from the dataset. It helps in increasing the efficiency of machine learning algorithms. In future the process, the feature engineering procedures for the customer churn dataset which helps this report to be all set for machine learning model.

TYPE OF FEATURES

This dataset contains both category and numerical values. Discrete and continuous numerical features were kept apart and were not able to find any timestamp features in this dataset.

NUMERIC FEATURES

Five Numerical factors are included in this dataset. Monthly charges, Tenure months, Total charges, Churn value and Churn score are some of the factors.

```
Num of Numerical Features : 5

['Tenure Months',
 'Monthly Charges',
 'Total Charges',
 'Churn Value',
 'Churn Score']
```

Fig: No of Numeric Features

CATEGORICAL FEATURES

There are fifteen categorical characteristics in the dataset which are City, Senior Citizen, Partner, Dependents, Multiple Lines of services, Internet Service, Online Security, Backup, Device Protection Service, Tech Support contract, Streaming TV, Movies, Contract, Billing Type and Payment Type are few of those in Categorical Features.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

```
Num of Categorical Features : 15

['City',
 'Senior Citizen',
 'Partner',
 'Dependents',
 'Multiple Lines',
 'Internet Service',
 'Online Security',
 'Online Backup',
 'Device Protection',
 'Tech Support',
 'Streaming TV',
 'Streaming Movies',
 'Contract',
 'Paperless Billing',
 'Payment Method']
```

Fig: No of Categorical Features

DISCRETE FEATURES

Churn Value is the only feature which comes under discrete features in the dataset. This feature has only two values 0 and 1.

```
Num of Discrete Features : 1

['Churn Value']
```

Fig: No of Discrete Feature

CONTINUOUS FEATURES

There are only four continuous factors in the dataset. These factors are Churn Score, Monthly Charges, Total Charges and Tenure Months.

```
Num of Continuous Features : 4

['Tenure Months', 'Monthly Charges', 'Total Charges', 'Churn Score']
```

Fig: No of Continuous Features

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

MULTICOLLINEARITY CHECK

Multicollinearity does not provide stable data which leads to inappropriate results during the evaluation of model parameters. The coefficients which play a vital role will not be interpreted in a prompt manner. The Variance Inflation Factor (VIF) was used to evaluate multicollinearity and the same was found in continuous features. As the Monthly Charges feature has a sturdy correlation with Total Charges, the same was neglected and the multicollinearity issue has been sorted out.

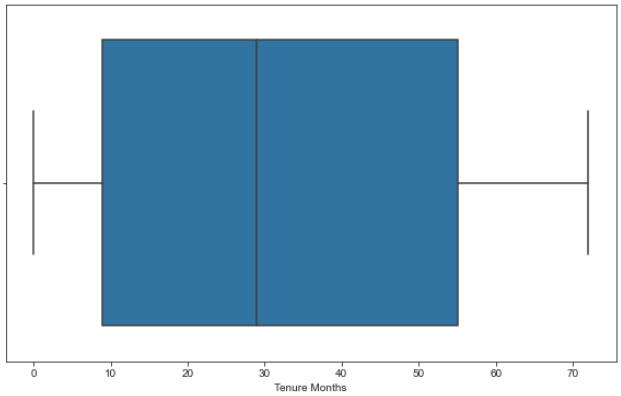
	Variable	VIF
0	Tenure Months	5.777478
1	Monthly Charges	3.308527
2	Total Charges	9.472717
3	Churn Score	1.106730

Fig: VIF Calculation

OUTLIER CHECK AND CAPPING

When evaluating a continuous feature for Outlier check, there were no Outlier found.

In the customer churn dataset, the engineering operations features are implemented such as determining the kind of features, examining multicollinearity and examining outliers. This has been done to make sure that the dataset is good to continue further for machine learning modelling.



Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

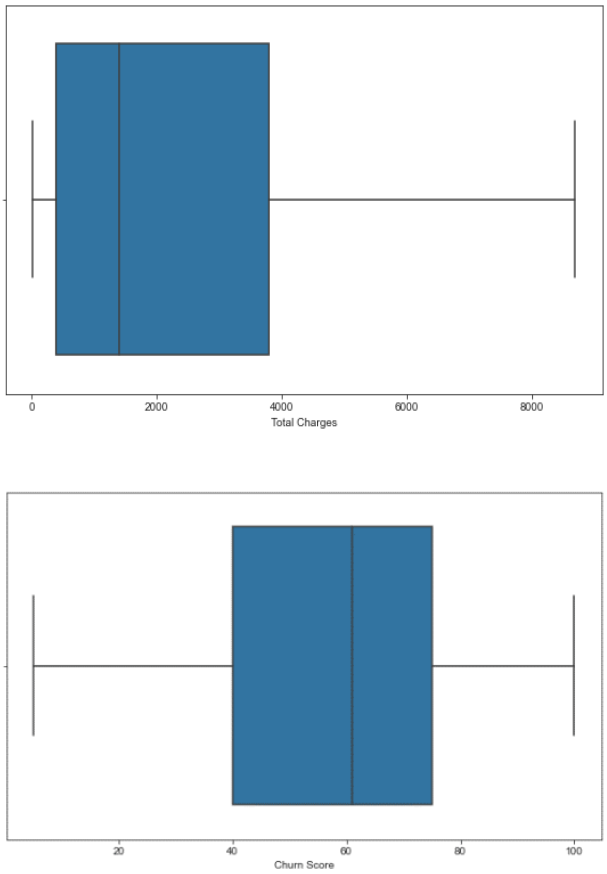


Fig: Checking the Outliers

FEATURE TRANSFORMATION

Feature transformation is always one of the mandatory steps in data preprocessing which helps in increasing the efficiency of machine learning algorithms. When starting a feature transformation, it is important to note that there may be an imbalance in the data so there are many strategies like power transformation which can be applied when addressing a skewed / imbalance in distribution of data.

To analyze the skewness in this dataset for continuous feature, skew() function can be used. As a result, the 'Total Charges' factor had an imbalance value of 0.96 which states that the data had a lack of balance. To solve this lacking, the power transformer was used which helps the dataset to spread equally without unevenness to avoid skewness.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

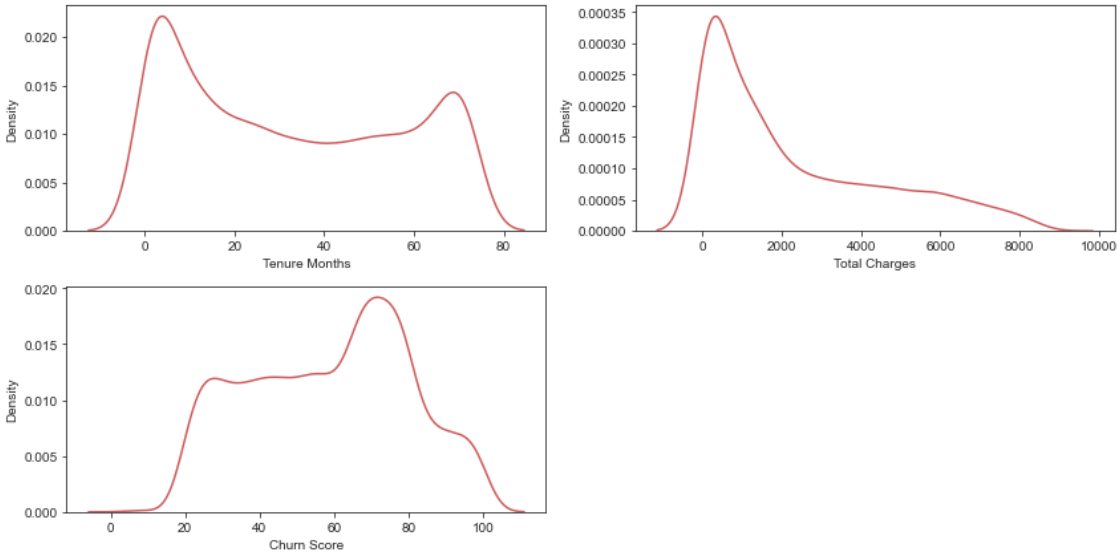


Fig: Checking Distribution of data for Feature Transformation

To show the distribution of data before and after scaling, the kdeplot() function has been used from the Seaborn library. The kernel density estimate (KDE) used above states that the feature of power transformed when used in "Total Charges" will make it more evenly distributed.

SPLIT X AND Y

In machine learning, it is mandatory to divide the datasets as X and Y. These datasets will be separated into X and Y variables. In difference to the other columns, where each factor is allocated to X variable, the Churn Value column alone has been allocated to Y variable because it is a dependent variable (Target Feature).

The number of distinct variables in each categorical feature has been given “for” loop in the code which aids in feature engineering. Finally, the dataset can be used now for machine learning algorithm analysis.

FEATURE ENCODING AND SCALING

one hot encoding for columns which had less unique values and not ordinal.

In the dataset, features encoding and scaling methods are used. On the columns where the features were not clear enough, one hot encoding is used to make sure it is clear enough to proceed further. Mean imputation has been done just to assure that the mean and median values are identical for continuous features. Features which have normal distribution of data have undergone processes such as standard scaling and the “Total Charges" feature has undergone power transformation. To create a column transformer, three different transformers such as power, category and numeric transformers are combined.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

It is mandatory to use ‘column transformer’ step in preprocessor data. It is used for applying standard scaling to numerical features, one hot encoding to categorical features, power transformation with imputation and standard scaling to the 'Total Charges' feature and standard scaling to the feature of power consumption. Fitting in the preprocessor to the original X matrix yields the final modified X matrix.

METHODOLOGY

HANDLING IMBALANCED DATASET

The telecom churn dataset is uneven which means that the number of churned customers is much lower than the quantity of non – churned customers. To investigate this crisis, the SMOTE Tomek was used to equalize the dataset. To create a balanced dataset, this technique helps in combining oversampling of the minority class with under sampling of the dominant class.

SMOTE-TOMEK

The SMOTE-Tomek method is a hybrid technique that combines the SMOTE oversampling method along with Tomek Links under sampling. The SMOTE- Tomek technique initially uses SMOTE to oversample the minority class and then it uses Tomek Links to eliminate the samples from majority class that are closely related to the minority class samples. As per the outcome, the SMOTE Tomek approach is useful for imbalanced datasets with a lower number of instances in the minority class when compared to the majority class.

TRAIN TEST SPLIT

The real dataset is separated into two different components in this technique, one is a training dataset and the other one is a testing dataset. The training dataset helps to train the machine learning model while the testing dataset is employed to evaluate the model’s performance with new or unknown data. The training dataset always shows a greater fraction of the original dataset when compared to the testing dataset which shows up a smaller portion. For instance, a common split uses 80% of the data for training and 20% for testing.

MODEL SELECTION

This Telecom dataset has been trained and evaluated with Logistic Regression, Gradient Boosting, CatBoosting Classifier, Random Forest, Decision Tree, K-Neighbors Classifier, XGBClassifier and AdaBoost Classifier. Below models have been used to compare the performance of various classification methods in a binary classification issue.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

LOGISTIC REGRESSION

In this project, the logistic regression machine learning algorithm was used to perform a binary classification on IBM Telecom dataset. It is a type of supervised Machine learning approach which is mostly used for binary classification when the target is to predict the probability based on the input features.

The reason for selecting this algorithm is that it is effective for binary classification. It works well when there is a linear relationship between input and target features and it makes it easy to understand the coefficients and strength between the features well.

The logistic regressions use the set of input features to estimate the target variable (Customer is likely to churn or not). It maps the input features with the target probability between 0 and 1. During the training, this algorithm fits the logistic function to train by perfecting the cost functions. The optimization process adjusts the parameters of the logistic function to reduce the difference between actual target values with predicted probabilities. Once the model is trained, it can be evaluated using the new instances.

```
Model : Logistic Regression
Model performance for Training set
- Accuracy: 0.9504
- F1 score: 0.9511
- Precision: 0.9383
- Recall: 0.9643
- Roc Auc Score: 0.9504
----------------------------------
Model performance for Test set
- Accuracy: 0.9070
- F1 score: 0.8471
- Precision: 0.7943
- Recall: 0.9075
- Roc Auc Score: 0.9072
==================================
```

Fig: Logistic Regression - Training and Testing Accuracy

The above result of Logistic Regression model has an accuracy in training is 0.95 and the accuracy in testing is 0.90 when feeding the new data into the model. The F1 score, Precision, Recall and Roc Auc Score has been calculated for both training and assessing the dataset for those which are above 0.90.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

GRADIENT BOOSTING

The Gradient boosting machine learning algorithm was used to perform a binary classification on the IBM Telecom dataset. It is a combination of different weak models like decision trees to create an effective strong model in the prediction.

The reason for selecting the model for this project is to perform a classification task which is specifically effective in handling complex and non-linear connections between the input parameters and the target feature. Additionally, this algorithm is immensely popular as it can work with high dimensional dataset because it is less prone to overfitting when compared with another tree-based algorithm. It works by iteratively adding each weak models (start with decision trees) to ensemble to fit the residual errors of each earlier model by updating the parameters of each weak models to reduce the difference between the targeted and actual values until it achieves the desire accuracy.

Here, the Gradient boosting model is used to predict if the customer's dropout or not and it is based on many features like demographics, services, and charges.

```
Model : Gradient Boosting
Model performance for Training set
- Accuracy: 0.9555
- F1 score: 0.9563
- Precision: 0.9399
- Recall: 0.9732
- Roc Auc Score: 0.9555
----------------------------------
Model performance for Test set
- Accuracy: 0.9134
- F1 score: 0.8565
- Precision: 0.8089
- Recall: 0.9100
- Roc Auc Score: 0.9124
==================================
```

Fig: Gradient Boosting - Training and Testing Accuracy

The above result of Gradient Boosting model has a Training accuracy of 0.95 and the Testing accuracy is 0.91 when feeding the new data into the model. The F1 score, Precision, Recall and Roc Auc Score has been calculated for both training and testing dataset which has a score of above 0.85.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

CATBOOSTING CLASSIFIER

The CatBoost classifier Machine learning algorithm was used to perform the Binary classification on IBM Telecom Dataset to predict customer attrition. This algorithm is specifically useful for high dimensional categorical data.

The reason for using this algorithm is that it is like a gradient boosting algorithm as it trains the weak models such as decision trees iteratively by adding new models to ensemble and fitting them with the residual error of previous models.

This algorithm is effective for handling categorical features without the requirement of extensive feature engineering. This algorithm is used to predict the customer attrition based on the available features in the dataset. This algorithm provides the best way to manage the high dimensional categorical features.

```
Model : CatBoosting Classifier
Model performance for Training set
- Accuracy: 0.9702
- F1 score: 0.9706
- Precision: 0.9575
- Recall: 0.9841
- Roc Auc Score: 0.9702
----------------------------------
Model performance for Test set
- Accuracy: 0.9141
- F1 score: 0.8551
- Precision: 0.8207
- Recall: 0.8925
- Roc Auc Score: 0.9076
==================================
```

Fig: CatBoosting Classifier - Training and Testing Accuracy

The above result of CatBoosting Classifier model has a accuracy in Training is 0.97 and the accuracy in Testing is 0.91 when feeding the new data into the model. The F1 score, Precision, Recall and Roc Auc Score has been calculated for both training and testing dataset which has a score of above 0.82.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

RANDOM FOREST

The Random Forest Algorithm was used to forecast the Customer turnover in the IBM Telecom dataset. This algorithm combines multiple decision trees to increase the accuracy and robustness of the model.

The reason for choosing this model is that it is powerful and versatile to oversee both numerical and categorical data. It is less prone to overfitting when compared to other algorithms. This algorithm is simple to implement and interpret which helps to find the most relevant variables to predict the outcome. It works by building a decision tree where each tree is trained with random subsets of training data with input features. The result is obtained by calculating the aggregate of each tree prediction / voting majority / weighted average.

The random forest is used to estimate the customer attrition in Telecom industry using various features.

```
Model : Random Forest
Model performance for Training set
- Accuracy: 1.0000
- F1 score: 1.0000
- Precision: 1.0000
- Recall: 1.0000
- Roc Auc Score: 1.0000
----------------------------------
Model performance for Test set
- Accuracy: 0.9077
- F1 score: 0.8338
- Precision: 0.8534
- Recall: 0.8150
- Roc Auc Score: 0.8797
==================================
```

Fig: Random Forest - Training and Testing Accuracy

The above result of Random Forest model has an accuracy in Training is 1.0 and the accuracy in Testing is 0.90 when feeding the new data into the model. The F1 score, Precision, Recall, Roc Auc Score has been calculated for both training and testing dataset which has a score of above 0.80.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

DECISION TREE

The Decision Tree was used on the IBM Telecom dataset to predict the churn status of the customer. This algorithm is powerful to learn the nonlinear boundaries by recursively partitioning the input space based on the input features value.

The reason for selecting this algorithm is that it is simple to interpret and visualize. It can also manage both numerical and categorical data. It keeps executing until it reaches the criterion such as maximum depth with minimum number of instances per node by splitting data into segments.

The Decision tree used in this project is to predict the customer churn which comes under binary classification and it provides insights about the importance of each instance to predict the outcome.

```
Model : Decision Tree
Model performance for Training set
- Accuracy: 1.0000
- F1 score: 1.0000
- Precision: 1.0000
- Recall: 1.0000
- Roc Auc Score: 1.0000
----------------------------------
Model performance for Test set
- Accuracy: 0.9113
- F1 score: 0.8459
- Precision: 0.8345
- Recall: 0.8575
- Roc Auc Score: 0.8951
==================================
```

Fig: Decision Tree - Training and Testing Accuracy

The above result of Decision Tree model has an accuracy in Training is 1.0 and the accuracy in Testing is 0.91 when analyzing the performance of the model. The F1 score, Precision, Recall, Roc Auc Score has been calculated for both training and testing dataset which has a score of above 0.84.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

K-NEIGHBORS CLASSIFIER

The K-Neighbors Classifier algorithm was used to perform the binary classification on the IBM dataset of Telecom Churn. This algorithm is effective to learn the nonlinear decision boundaries by finding the K- nearest instances based on the distance metric in the trainset and then classify the new instances by majority vote of its neighbors.

The reason for selecting this algorithm is that it is a non-parametric algorithm which does not make any assumptions about the underlying distribution of the data and it can handle both numerical and categorical data. This algorithm can be used in imbalanced dataset. When training an IBM Telecom dataset with this model, it stores the training instances and class labels in the memory.

K-Neighbors classifier is used to predict whether the customer churns or not which is based on numerous factors. This algorithm is simple and effective.

```
Model : K-Neighbors Classifier
Model performance for Training set
- Accuracy: 0.9220
- F1 score: 0.9274
- Precision: 0.8676
- Recall: 0.9959
- Roc Auc Score: 0.9220
----------------------------------
Model performance for Test set
- Accuracy: 0.8325
- F1 score: 0.7602
- Precision: 0.6404
- Recall: 0.9350
- Roc Auc Score: 0.8634
==================================
```

Fig: K-Neighbors Classifier - Training and Testing Accuracy

The above result of K-Neighbors Classifier model has an accuracy in Training is 0.92 and the accuracy in Testing is 0.83 when analyzing the performance of the model. The F1 score, Precision, Recall and Roc Auc Score has been calculated for both training and testing dataset for values which are above 0.64.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

XGB CLASSIFIER

The XGBoost (Extreme Gradient Boosting) Classifier algorithm is used to perform a Binary classification on the IBM Telecom dataset. This algorithm is a popular ensemble algorithm which combines the strength of both Gradient Boosting and Random Forest.

The reason for choosing the algorithm for this project is, it is known for high accuracy and scalability. It can also manage both numerical and categorical data. It can predict the complex relationship between the input and target variables. Although this algorithm is computationally expensive.

This is a tree-based gradient boosting method that outperforms classic gradient boosting in terms of speed and execution time.

```
Model : XGBClassifier
Model performance for Training set
- Accuracy: 0.9925
- F1 score: 0.9925
- Precision: 0.9899
- Recall: 0.9952
- Roc Auc Score: 0.9925
----------------------------------
Model performance for Test set
- Accuracy: 0.9233
- F1 score: 0.8670
- Precision: 0.8544
- Recall: 0.8800
- Roc Auc Score: 0.9103
==================================
```

Fig: XGB Classifier - Training and Testing Accuracy

The above result in the screenshot shows that the XGBoost Classifier algorithm has been trained with the IBM Telecom dataset and it has a Training accuracy of 0.99 and the Testing accuracy is 0.92 when analyzing the performance of the model. The F1 score, Precision, Recall and Roc Auc Score has been calculated for both training and testing dataset which has a score of above 0.85.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

ADABOOST CLASSIFIER

The AdaBoost Classifier (Adaptive Boosting) algorithm is used to perform the Binary classification using IBM Telecom dataset. This algorithm creates a strong classifier by combining the strengths of multiple weak learners.

The reason for using this algorithm is that it is good for high accuracy and robustness to noise in the data. It can also handle both numerical and categorical data. It works by fitting the weak classifier iteratively. During the training, this algorithm assigns a more weight to the misclassified instances and less weight to the properly classified instances and then trains the next weak classifier on the reweighted data.

The outcome of this algorithm is the average prediction of all the weak classifiers with the higher weights assigned to the more exact classifier. This project used IBM Telecom Customer dataset to this model to check whether the customer will stay or not.

```
Model : AdaBoost Classifier
Model performance for Training set
- Accuracy: 0.9526
- F1 score: 0.9533
- Precision: 0.9406
- Recall: 0.9662
- Roc Auc Score: 0.9526
----------------------------------
Model performance for Test set
- Accuracy: 0.9184
- F1 score: 0.8639
- Precision: 0.8202
- Recall: 0.9125
- Roc Auc Score: 0.9166
==================================
```

Fig: AdaBoost Classifier - Training and Testing Accuracy

This is an algorithm that combines several weak learners models (often decision trees) to produce a strong segregator. The above result in the screenshot shows that the AdaBoost Classifier algorithm has been trained with IBM Telecom dataset and its Training accuracy is 0.95 and the Testing accuracy of 0.91 when analyzing the performance of the model. The F1 score, Precision, Recall, Roc Auc Score has been calculated for both training and testing dataset which has a value of over 0.82.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

MODEL TRAINING AND EVALUATION

To neglect the biased performance measurements and data leaks, a pipeline technique is used in which SMOTE is applied only to the training after train test split. The SMOTE Tomek technique is used with a random seed of forty-two and a minor sampling strategy.

Each model's performance was assessed using accuracy, F1 score, precision, recall, and ROC AUC score.

	Model Name	Accuracy	F1 Score	Precision	Recall	ROC AUC Score	Execution Time
0	XGBClassifier	0.923350	0.866995	0.854369	0.8800	0.910268	0.786592
1	AdaBoost Classifier	0.918382	0.863905	0.820225	0.9125	0.916607	1.357430
2	CatBoosting Classifier	0.914123	0.855090	0.820690	0.8925	0.907598	20.114514
3	Gradient Boosting	0.913414	0.856471	0.808889	0.9100	0.912384	5.674489
4	Decision Tree	0.911285	0.845869	0.834550	0.8575	0.895053	0.310061
5	Random Forest	0.907736	0.833760	0.853403	0.8150	0.879750	10.121411
6	Logistic Regression	0.907026	0.847141	0.794311	0.9075	0.907169	0.258730
7	K-Neighbors Classifier	0.832505	0.760163	0.640411	0.9350	0.863437	0.006012

Fig: Evaluation Metrics of all Model

Performance metrics such as Accuracy, F1 score, Recall and ROC Auc are used to assess the performance of these models. Accuracy is the proportion of correct forecasts to overall predictions. The harmonic mean of precision, recall is the F1 Score which provides the result with the combination of precision and recall which is effective.

Precision is a ratio of true positive forecasts to overall positive predictions. Recall is the fraction of true positive predictions in the data over the total number of positive cases. ROC Auc score is the classifier performance across the various threshold.

The gradient Boosting model performance is good when compared with the other implemented model performance for this dataset. The accuracy of this model is 0.91, F1 Score is 0.85 and ROC AUS Score is 0.91. The XGBClassfier nearly followed the model with an accuracy of 0.92, Precision of 0.86 and the ROC AUC Score is 0.91.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

HYPER PARAMETER TUNING

To improve the machine learning model performance effectively, it is mandatory to have Tuning hyperparameters which helps in increasing the efficiency of the model. Hyperparameters help in providing approaches to the problem while creating the algorithms which in turn gives the perfect outcome without many complications. Each algorithm has different hyperparameters and they are not ideal for every dataset. As an outcome, it is critical to tweak the hyperparameters to maximize the algorithm's performance for a certain dataset.

Even though many algorithms have been implemented, the hyperparameter is mandatory because it plays vital role in adjustments which helps in giving the perfect results. After trying various approaches, Gradient Boosting approach performs best on the dataset. However, by utilizing hyperparameter tuning for Gradient Boosting, XGB Classifier and AdaBoost Classifier methods which can be tuned in a prompt way to reach even greater performance. The act of selecting the ideal values for model parameters that were not learned during the initial training process is known as hyperparameter tuning.

RETRAINING THE MODEL WITH BEST PARAMETERS

On the dataset, three models were trained and estimated: Gradient Boosting, XGBClassifier and AdaBoost Classifier. Tuned hyperparameters were used to train all models. According to the results, all three models perform equally with accuracy and F1 scores ranging from 0.909 to 0.920.

```
Model : XGBClassifier
Model performance for Training set
- Accuracy: 0.9619
- F1 score: 0.9624
- Precision: 0.9489
- Recall: 0.9764
- Roc Auc Score: 0.9619
----------------------------------
Model performance for Test set
- Accuracy: 0.9198
- F1 score: 0.8656
- Precision: 0.8254
- Recall: 0.9100
- Roc Auc Score: 0.9168
==================================
```

Fig: Performance of the XGBClassfier model after Parameter Tuning

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

When dealing with an uneven dataset, pay special attention to the recall metric which calculates the model's ability to correctly detect positive cases (in this example, churning consumers). As per the outcome, the model with the highest recall score should be prioritized.

The tuned Hyperparameter was used to train all these models. According to the result, these three models perform similarly with F1 Scores ranging from 0.90 to 0.92 for accuracy. The recall statistic which assesses how well a model can identify positive cases (Churn customer) needs extra care while dealing with an unbalanced dataset. As a result, consider the model with the highest recall score priority.

```
Model : AdaBoost Classifier
Model performance for Training set
- Accuracy: 0.9361
- F1 score: 0.9374
- Precision: 0.9191
- Recall: 0.9564
- Roc Auc Score: 0.9361
----------------------------------
Model performance for Test set
- Accuracy: 0.9155
- F1 score: 0.8621
- Precision: 0.8035
- Recall: 0.9300
- Roc Auc Score: 0.9199
==================================
```

Fig: Performance of the AdaBoost Classifier model after Parameter Tuning

The AdaBoost Classifier with a Recall score of 0.93 which is close to XGBClassfier has the highest Recall of 0.91. It is crucial to choose a model based on interpretability, execution time and amount of computational power needed. The XGBClassifier has the greatest ROC AUS score of 0.91 which shows that the overall performance is good based on the rating of the positive and negative classes. An effective tree-based model implementation of the gradient boosting algorithm is used in XGBClassifier.

The Recall score of the Gradient Boosting Model is 0.88. So, as a conclusion of completing the best model is XGBClassfier for this dataset which is based on the evaluation metrics. The supervised learning technique is applied to both classification and regression issues. The approach builds a group of poor decision tree models and then enhances their performance by changing the weight of incorrectly classified data throughout each iteration.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

```
Model : Gradient Boosting
Model performance for Training set
- Accuracy: 1.0000
- F1 score: 1.0000
- Precision: 1.0000
- Recall: 1.0000
- Roc Auc Score: 1.0000
----------------------------------
Model performance for Test set
- Accuracy: 0.9191
- F1 score: 0.8613
- Precision: 0.8389
- Recall: 0.8850
- Roc Auc Score: 0.9088
==================================
```

Fig: Performance of the Gradient model after Parameter Tuning

MODEL PERFORMANCE EVALUATION AFTER PARAMETER TURNING

As per the outcome, the original Gradient Boosting model achieved the greatest ROC AUC score of 0.90. After dealing with the dynamic data and fine-tuning the hyperparameters, the XGBClassifier model secured the highest ROC AUC score of 0.96 on the train set and 0.91 on the new data (test set). In terms of overall evaluation, this suggests that the XGBClassifier model outperforms when compared to other models.

	Model Name	Accuracy	F1 Score	Precision	Recall	ROC AUC Score	Execution Time
0	XGBClassifier	0.919801	0.865636	0.825397	0.910	0.916843	1.050378
1	Gradient Boosting	0.919092	0.861314	0.838863	0.885	0.908803	8.991202
2	AdaBoost Classifier	0.915543	0.862109	0.803456	0.930	0.919906	0.110843

Fig: Performance Evaluation after Parameter Turning

Observing the process after dealing with uneven data and changing the hyperparameters, the performance of all the three models increased. This states that the necessity of dealing with uneven data before model training plays a vital role.

According to the Recall measure, ROC AUC Score and the overall performance, XGBClassifier is the best model for providing the information on Customer attrition in this dataset.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

WITH IMBALANCED DATASET

Based on ratings, the Month-to-month contract and Churn Score are the most influential features for the expected results. Customers with a month-to-month contract and a high churn score are more likely to churn according to this data. Other features such as internet service type, online security, payment method and the customers who have dependents have high relevance score.

Feature important scores can be skewed while dealing with uneven datasets and this is the crucial point to be noted. In this scenario, the number of churners is lower when compared to the number of non-churners in uneven dataset. As an outcome, the model may be biased towards the non-churned customers (majority class) and may not place much emphasis on factors that are more relevant for predicting the churned customers (minority class).

CHECKING PERFORMANCE OF ML ALGORITHM FOR IMBALANCED DATASET

As per the outcome, the most influential feature for predicting attrition is Month-to-month contract followed by attrition Score, number of dependents, Internet Service and Fiber optic. The model can be increased more by changing hyperparameters or using various techniques. However, the given imbalanced dataset to the existing model works quite well while performing the classification on biased dataset to forecast the customer turnover. The model uses the class weights to accommodate the imbalance. These findings can be improved in the earlier section by handling imbalance through SMOTE technique which makes the model work effectively.

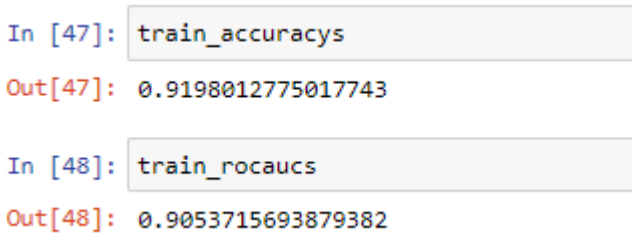


Fig: Performance of XGBoost for Imbalanced Dataset

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

SHAP – SHAPley Additive exPlanations

The SHAP values define the contribution of each feature to the final prediction and those are all used to describe the output of machine learning methods. Using the Shap_Explainer() function which generates an explainer object for an XGBClassifier model with the features names provided. Then it uses the explanation object to compute the SHAP values for a single instance of test data. For example, randomly checking the result for the rows in the test data as sample input of all features and then sets the results to the variable shap_values.

Finally, it shows in the result a waterfall plot with the stored shap_values for the instances of the test data using the shap.plots.waterfall() function. The below plot depicts each feature's contribution to the final forecast. The positive values stand for the features which raise the prediction and Negative values which indicates the feature that decreases the predictions.

In the below plot, the boxes with blue colour reflect factors which drive the model prediction towards lower values while the red colour describes features that decrease the model prediction towards higher values in the SHAP waterfall plot. The features "churn_score" and "total charges" are shown in red colour with a positive SHAP value, states that the high value of this feature contributes positively to the model's prediction.

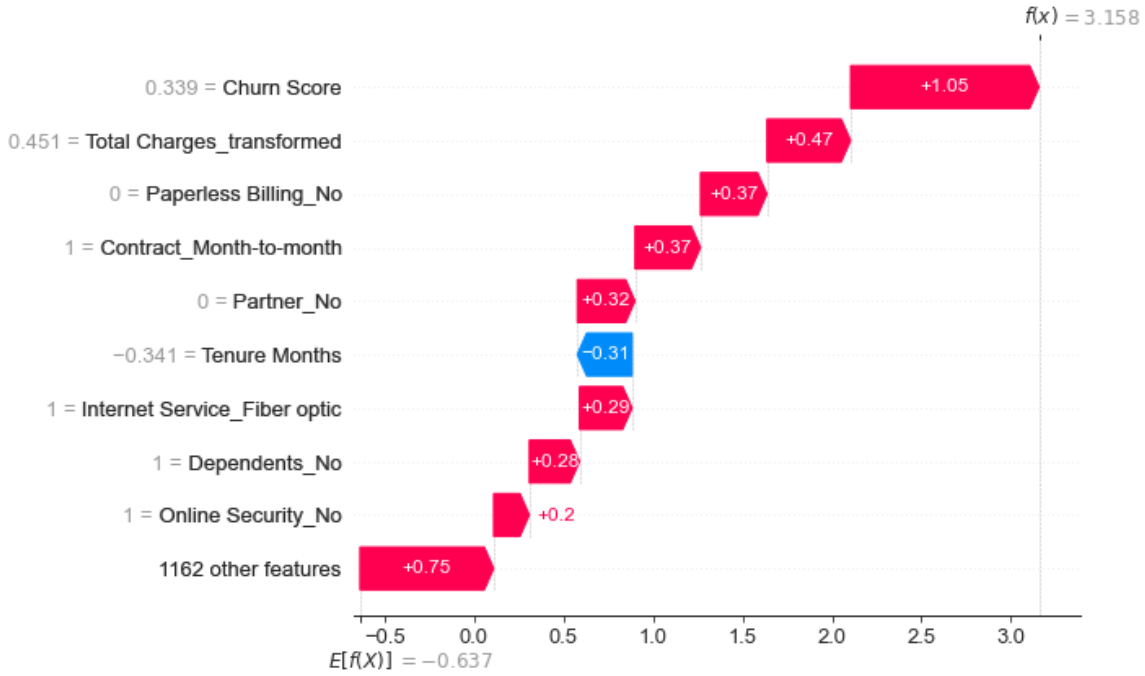


Fig: SHAP result for sample inputs

In the above sample input SHAP waterfall plot, the feature "Tenure Months" is described in blue color with a negative SHAP value (-0.31) which states that the feature has a negative impact on the model's forecasts. According to this input, the Feature Tenure months affect the customer to continue with the company.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

SHAP WITH XGBOOST MODEL

The XGBoost model finds the top two characteristics which may enhance the chance of churn for a given customer demographics and service types with charges on the SHAP study. To create the prediction model, it is good to consider all the features such as Demographics (Age, Gender, Partner, Dependents), Service information (Contract type, Internet Service, Online back, Tech support, etc) with Charges (Monthly charges, Total charges). Mostly, the top two features that may enhance the customer's churn are based on the inputs of each feature. These top two suggestions are effective for businesses to hold the customer by providing offers to boost client retention. This method is effective to identify the at-risk customers who need extra care and support with service satisfaction from the Telecom Industry. As a result, this model predicts and prints results based on the input values.

SHAP RESULT

The XGBoost model was evaluated with sample input information from a client (End user) who lives in Los Angeles and they do not have partners or dependents. The customer is not elderly and he has been the only customer for two months. The customer has DSL internet access and holds online security and backup but does not have device protection or technical support. The consumer does not watch streaming TV or movies online, month-to-month contract type, receives paperless billing and the payments are made using postal checks. The consumer owes 108.15 in total charges and has a churn score of 86. The model forecasts that the customer would churn (y_pred=1) and generate feedback based on the top two factors that trigger this likelihood. As per the feedback, the attrition score and tenure months are the top two factors that may increase the chance of attrition.

It is vital to remember that the churn score and tenure months contribute positively to the risk of churning does not mean that the customer churns. Based on the model analysis in the Telecom dataset, customers with higher churn rate and short tenure months are more likely to churn.

By using this waterfall SHAP, its effectively found on how each feature affects the churner rate in the Telecom Industry. It is the best solution for the industry to concentrate on keep the customers by identify their dissatisfaction.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

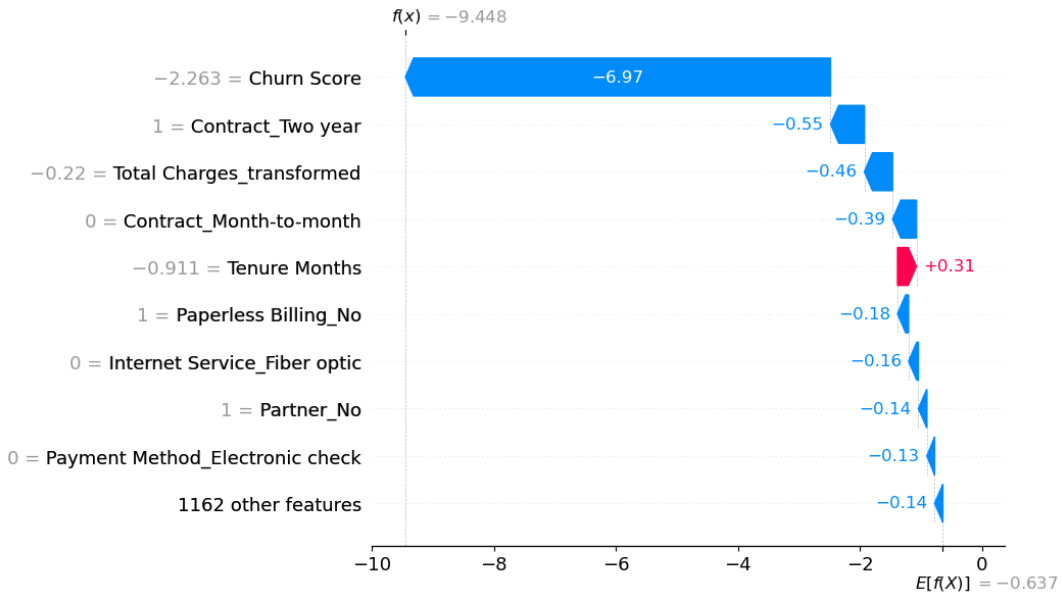


Fig: Prediction of Telecom Customer Attrition

FASTAPI WEB APPLICATION

After considering some inputs, a FastAPI web application is developed to forecast if the customer is like to churn or not. The model computes SHAP values for the input data to decide which factors are most relevant in forecasting the attrition rates. When the final user (Telecom industry) submits the input form, the program uses the preprocessor object to process the input data, the XGBoost model is used to predict the SHAP values and the SHAP library is used to create a waterfall plot of the SHAP values.

If the prediction is 1 (high churn likelihood), the program calculates the exact SHAP values in descending order, generates the names of the top two factors with the highest SHAP values and offers a feedback message for the user says which factor may raise the churn likelihood. The program also saves the created plot to a static file and sends the user an HTML response with the forecasted result.

Below attached screenshot is the end user Web App which has the index page to get the end user to enter the customer details and check important which affecting the customer to churn.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

CHURN PREDICTION

City

Fresno

Senior Citizen

Yes

Partner

Yes

Dependents

No

Multiple Lines

No

Internet Service

DSL

Online Security

No

Online Backup

Yes

Device Protection

No

Tech Support

Yes

Streaming TV

No

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

Streaming Movies

Yes

Contract

Two year

Paperless Billing

Yes

Payment Method

Bank transfer (automatic)

Tenure Months [0-75]

13

Total Charges [12-9000]

1800

Churn Score [0-100]

50

Predict

Prediction : Low Chance of Churn

Feedback : Based on the provided information, we do not see any major issues that would result in a high likelihood of churn.

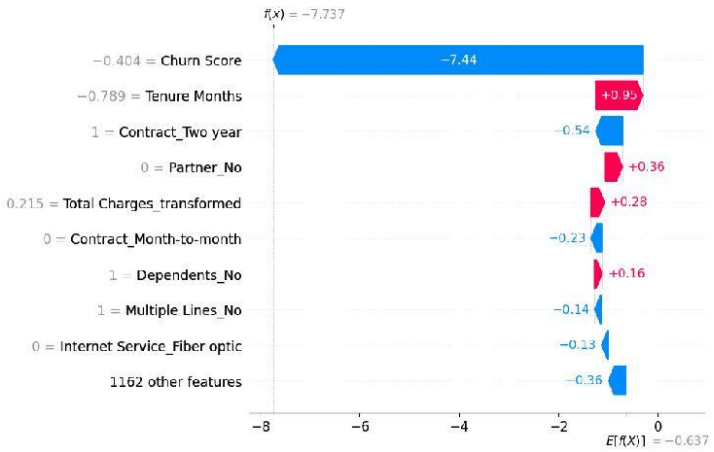


Fig: FastAPI web application with result

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

RESULT

The SHAP technique was used in this project to explain the predictions of the XGBoost classifier model which makes easier for the Telecom Industry (Final user) to interact with the model and obtain the SHAP feedback about the most affecting features.

This web application uses Fast API built for the user to input the services and details of customer to get feedback on the top two features which influence the likelihood of churn. The SHAP waterfall plot provides the contribution of each feature to the result. The FastAPI uses the finalized XGBoost Classifier model to make prediction on input data by applying SHAP techniques that makes the user (Telecom Industry) understand the feedback and to take actionable retention.

ETHICAL CONCERNS

Any project which involves gathering and analyzing information from individuals will always face some kind of ethical issues. Below mentioned are some ethical issues in the project:

Data privacy: It is quite difficult to ensure that the information gathered from users is kept secure as it is available publicly on IBM blog site. This includes taking measures to protect data and prevent unauthorized access. But it will not be a majorly considerable ethical concern as it is a fictional dataset and there is no specific personal information such as name and phone number.

Bias: Biases in the data which is used to train machine learning algorithms can be perpetuated. It is difficult to train the model on diverse data and to look after its performance for any biases that may create.

Consent: It is mandatory to inform the users about the information which is collected and let them know how it is going to be used in the future. Based on their consent the details can be used further with the data when this project is launched in real time. But, in this study it is a fictional dataset provided by IBM.

Transparency: It is difficult to be open about how the model works, what data it relies on and how it makes judgements. This can aid users in understanding how the collected information and model may affect them in making informed decisions about whether to share the information or not.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

DISCUSSION AND CONCLUSION

In this project, tools like XGBoost and SHAP have been used to create an attrition prediction model for a telecom company. To train this model, datasets played a vital role which included consumer demographics, service usage and account information. On this test set, there was an accuracy of 92% after preprocessing and feature engineering. Then, using FastAPI, a web application is built and deployed with Unicorn. Users (Telecom Industry) can enter the information and receive a prediction of whether they are at a high or low risk of churning. This project also provides feedback on the top two vital reasons that may increase the chance of attrition and offers steps to improve client retention.

One potential restriction of this project is its reliance on the information provided by the users. The dataset may not be representative of all customers and other factors influencing attrition are not included in the dataset which may exist. Furthermore, if client behaviors and preferences are dynamic then the model may become less accurate. This model could be used by the organization to target high-risk consumers with aggressive marketing strategies or to maximize the customer retention efforts for high-value customers. This could result in the treatment of specific consumers and to save the reputation of the company.

Finally, the study shows how to utilize machine learning to predict churn and gives the ‘web application’ for telecom industries to enhance customer retention. However, it is difficult to calculate the approach of potential limitations.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

FUTURE WORK

Here are some great ideas for future project work:

Include more features: Currently, the model forecasts attrition only using selective features. More relevant features must be added to the model in the future to increase its accuracy.

Investigate different algorithms: While XGBoost is a powerful algorithm, it may not always be the best choice in every circumstance. Many algorithms must be included to get accurate results when compared to XGBoost.

Deploy the model to production: The current project is offering a web interface through which users can enter the input data and estimate the customer churn. The model might be implemented in a production environment in the future such as a serverless function or a containerized application to give more scalable and dependable solutions.

Consider other Sampling methods: The current dataset is imbalanced with lower cases of Churner rate. The SMOTE technique has been used to balance the distribution of data. There is few advanced Machine learning that has the capability of balancing the dataset and perform effectively. Those can be implemented in future work of this project and different sampling methods can also be used.

Conduct user research: In this project, the feedback offered to users is based on the model combination of the input data. However, conducting user research to calculate whether the feedback is useful, actionable and relevant to their requirements would be much beneficial. This could help to increase the overall user experience of the application.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

REFERENCES

[1] STAMFORD, Conn., April 28, 2022, Gartner Says U.S. Total Annual Employee Turnover Will Likely Jump by Nearly 20% From the Prepandemic Annual Average.

[2] Telecom Services Market Size, Share & Trends Analysis Report by Service Type (Mobile Data Services, Machine-To-Machine Services), By Transmission (Wireline, Wireless), By End-use, By Region, And Segment Forecasts, 2023 - 2030 Historical Range: 2017 – 2021.

[3] Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms May 2020. International Journal of Engineering and Technical Research V9(05). Authors: V. Kavitha G. Hemanth Kumar S. V Mohan Kumar M. Harish.

[4] Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform. Authors: Muhammad Joolfoo University of Mauritius Rameshwar A Jugurnauth Khalid Joolfoo University of Mauritius.

[5] A comparison of machine learning techniques for customer churn prediction Author links open overlay panelT. Vafeiadis a, K.I. Diamantaras b, G. Sarigiannidis a, K.Ch. Chatzisavvas.

[6] A review and analysis of churn prediction methods for customer retention in telecom industries Publisher: IEEE. Ammara Ahmed; D. Maheswari Linen All Authors.

[7] Customer churn prediction system: a machine learning approach Springer February 2022. Computing. Authors: Praveen Lalwani Indian Institute of Technology (ISM) Dhanbad Manas Kumar Mishra VIT Bhopal University Jasroop Singh Chadha VIT University Pratyush Sethi VIT University.

[8] Machine-Learning Techniques for Customer Retention: A Comparative Study Sahar F. Sabbeh Faculty of computing and information sciences, King AbdulAziz University, KSA Faculty of computing and information sciences, Banha University, Egypt.

[9] A new feature set with new window techniques for customer churn prediction in land-line telecommunications Author links open overlay panelB.Q. Huang a, T.-M. Kechadi a, B. Buckley b, G. Kiernan b, E. Keogh b, T. Rashid a.

[10] Churn Prediction using Neural Network based Individual and Ensemble Models Publisher: IEEE. Mehpara Saghir; Zeenat Bibi; Saba Bashir; Farhan Hassan Khan All Authors.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

[11] A Customer Churn Prediction Model in Telecom Industry Using Boosting Publisher: IEEE. Ning Lu; Hua Lin; Jie Lu; Guangquan Zhang All Authors.

[12] Modeling customer churn in a non-contractual setting: the case of telecommunications service providers Ali Tamaddoni Jahromi,Mohammad Mehdi Sepehri ,Babak Teimourpour &Sarvenaz Choobdar . 27 Sep 2010, Published online: 13 Dec 2010.

[13] A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA Maha Alkhayrat, Mohamad Aljnidi & Kadan Aljoumaa Journal of Big Data volume 7, Article number: 9 (2020).

[14] Analysis of customer churn prediction in telecom industry using decision trees and logistic regression Publisher: IEEE. Preeti K. Dalvi; Siddhi K. Khandge; Ashish Deomore; Aditya Bankar; V. A. Kanade All Authors.

[15] Customer churn analysis in telecom industry Publisher: IEEE. Kiran Dahiya; Surbhi Bhatia All Authors.

[16] A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector Publisher: IEEE. Irfan Ullah; Basit Raza; Ahmad Kamran Malik; Muhammad Imran; Saif Ul Islam; Sung Won Kim All Authors.

[17] An Empirical Study on Customer Churn Behaviours Prediction Using Arabic Twitter Mining Approach by Latifah Almuqren, Fatma S. Alrayes 1 andAlexandra I. Cristea.

[18] Telco customer churn (11.1.3+) By Samples Team posted Thu July 11, 2019, 08:15 AM In IBM Cognos Analytics 11.1.3. The Telco customer churn data has information about a fictional telco company that supplied home phone and Internet services to 7043 customers in California in Q3.

[19] Churn prediction methodologies in the telecommunications sector: A survey Publisher: IEEE. W. M. C. Bandara; A. S. Perera; D. Alahakoon All Authors.

[20] CHURN PREDICTION TECHNIQUES IN TELECOM INDUSTRY FOR CUSTOMER RETENTION: A SURVEY Anam Bansal Assistant Professor, Computer Science and Technology, Central University of Punjab, Bathinda, Punjab, India.

[21] Machine learning based customer churn prediction in home appliance rental business Youngjung Suh Journal of Big Data volume 10, Article number: 41 (2023).

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

[22] Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study Authors: Wissam Nazeer Wassouf, Ramez Alkhatib, Kamal Salloum, Shadi Balloul Published in: Journal of Big Data | Issue 1/2020.

[23] Customer churn analysis for a software-as-a-service company April 2017 DOI:10.1109/SIEDS.2017.7937698 Conference: 2017 Systems and Information Engineering Design Symposium (SIEDS).

[24] Exploring consumer adoption of new services by analyzing the behavior of 3G subscribers: An empirical case study March 2012 Electronic Commerce Research and Applications 11(2). Authors: Li Chen Cheng National Taipei University of Technology Li-Min Sun.

[25] Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study Wissam Nazeer Wassouf, Ramez Alkhatib, Kamal Salloum & Shadi Balloul.

[26] Churn Prediction in Telecommunication Industry using Decision Tree. Authors: Nisha Saini, Monika, Dr. Kanwal Garg.

[27] Customer segmentate Authors: tegy development based on customer lifetime value: A case study Author links open overlay panelSu-Yeon Kim a, Tae-Soo Jung b, Eui-Ho Suh c, Hyun-Seok Hwang d.

[28] Framework for Targeting High Value Customers and Potential Churn Customers in Telecom using Big Data Analytics January 2017International Journal of Education and Management Engineering. Authors: Inderpreet Singh Simon Fraser University Sukhpal Singh Gill Queen Mary, University of London

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

APPENDIX

Abbreviations:

ML – Machine Learning
SVM - Support Vector Machine
PCA - Principal Component Analysis
RNN - Recurrent Neural Networks
GBM - Gradient Boosting Machine
IQR - Interquartile Range
CLTV - Customer LifeTime Values
VIF - Variance Inflation Factor
KDE - Kernel Density Estimate
XGBoost - Extreme Gradient Boosting
AdaBoost Classifier - Adaptive Boosting
SHAP – SHAPley Additive exPlanations

The IBM sample Telecom Customer Churn Dataset:

	CustomerID	Count	Country	State	City	Zip Code	Lat Long	Latitude	Longitude	Gender	...	Contract	Paperless Billing	Payment Method	Monthly Charges	Total Charges
0	3668-QPYBK	1	United States	California	Los Angeles	90003	33.964131, -118.272783	33.964131	-118.272783	Male	...	Month-to-month	Yes	Mailed check	53.85	108.15
1	9237-HQITU	1	United States	California	Los Angeles	90005	34.059281, -118.30742	34.059281	-118.307420	Female	...	Month-to-month	Yes	Electronic check	70.70	151.65
2	9305-CDSKC	1	United States	California	Los Angeles	90006	34.048013, -118.293953	34.048013	-118.293953	Female	...	Month-to-month	Yes	Electronic check	99.65	820.5
3	7892-POOKP	1	United States	California	Los Angeles	90010	34.062125, -118.315709	34.062125	-118.315709	Female	...	Month-to-month	Yes	Electronic check	104.80	3046.05
4	0280-XJGEX	1	United States	California	Los Angeles	90015	34.039224, -118.266293	34.039224	-118.266293	Male	...	Month-to-month	Yes	Bank transfer (automatic)	103.70	5036.3

5 rows × 33 columns

	City	Zip Code	Lat Long	Latitude	Longitude	Gender	...	Contract	Paperless Billing	Payment Method	Monthly Charges	Total Charges	Churn Label	Churn Value	Churn Score	CLTV	Churn Reason
	Los Angeles	90003	33.964131, -118.272783	33.964131	-118.272783	Male	...	Month-to-month	Yes	Mailed check	53.85	108.15	Yes	1	86	3239	Competitor made better offer
	Los Angeles	90005	34.059281, -118.30742	34.059281	-118.307420	Female	...	Month-to-month	Yes	Electronic check	70.70	151.65	Yes	1	67	2701	Moved
	Los Angeles	90006	34.048013, -118.293953	34.048013	-118.293953	Female	...	Month-to-month	Yes	Electronic check	99.65	820.5	Yes	1	86	5372	Moved
	Los Angeles	90010	34.062125, -118.315709	34.062125	-118.315709	Female	...	Month-to-month	Yes	Electronic check	104.80	3046.05	Yes	1	84	5003	Moved
	Los Angeles	90015	34.039224, -118.266293	34.039224	-118.266293	Male	...	Month-to-month	Yes	Bank transfer (automatic)	103.70	5036.3	Yes	1	89	5340	Competitor had better devices

Fig: The dataset

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

The Dataset Columns Description:

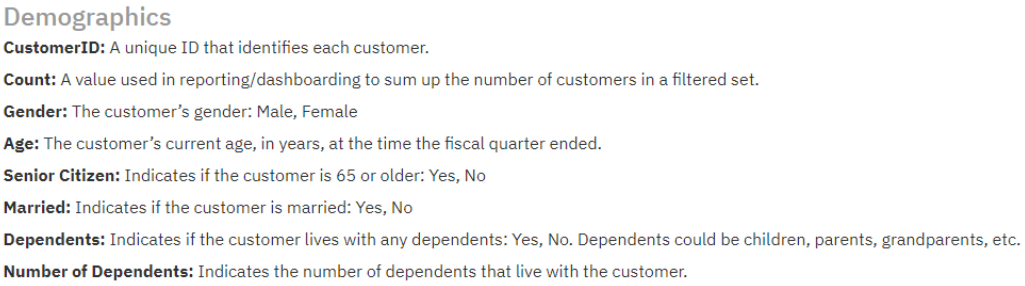


Fig: Customer Demographics details in the dataset

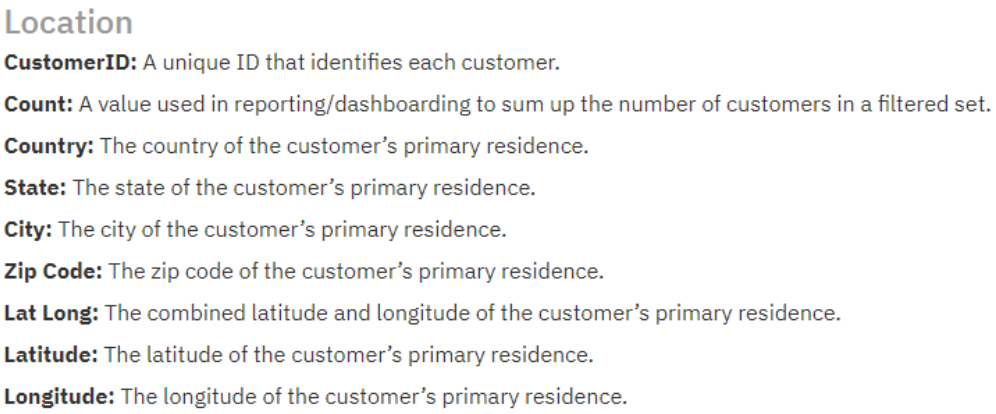


Fig: Customer Location details in the dataset

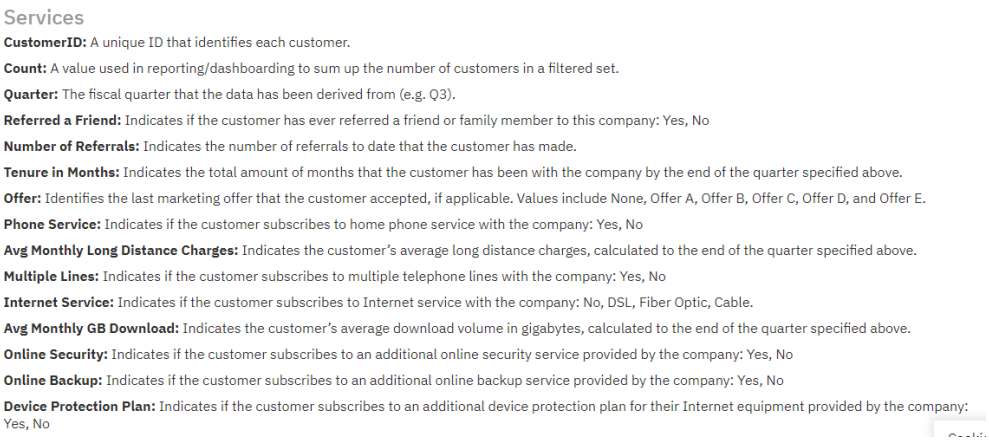


Fig: Service details in the dataset

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

Premium Tech Support: Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times: Yes, No

Streaming TV: Indicates if the customer uses their Internet service to stream television programing from a third party provider: Yes, No. The company does not charge an additional fee for this service.

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.

Streaming Music: Indicates if the customer uses their Internet service to stream music from a third party provider: Yes, No. The company does not charge an additional fee for this service.

Unlimited Data: Indicates if the customer has paid an additional monthly fee to have unlimited data downloads/uploads: Yes, No

Contract: Indicates the customer's current contract type: Month-to-Month, One Year, Two Year.

Paperless Billing: Indicates if the customer has chosen paperless billing: Yes, No

Payment Method: Indicates how the customer pays their bill: Bank Withdrawal, Credit Card, Mailed Check

Monthly Charge: Indicates the customer's current total monthly charge for all their services from the company.

Total Charges: Indicates the customer's total charges, calculated to the end of the quarter specified above.

Fig: Service Details in the dataset

Churn Label: Yes = the customer left the company this quarter. No = the customer remained with the company. Directly related to Churn Value.

Churn Value: 1 = the customer left the company this quarter. 0 = the customer remained with the company. Directly related to Churn Label.

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.

Churn Score Category: A calculation that assigns a Churn Score to one of the following categories: 0-10, 11-20, 21-30, 31-40, 41-50, 51-60, 61-70, 71-80, 81-90, and 91-100

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.

CLTV Category: A calculation that assigns a CLTV value to one of the following categories: 2000-2500, 2501-3000, 3001-3500, 3501-4000, 4001-4500, 4501-5000, 5001-5500, 5501-6000, 6001-6500, and 6501-7000.

Churn Category: A high-level category for the customer's reason for churning: Attitude, Competitor, Dissatisfaction, Other, Price. When they leave the company, all customers are asked about their reasons for leaving. Directly related to Churn Reason.

Churn Reason: A customer's specific reason for leaving the company. Directly related to Churn Category.

Fig: Churn Details in the dataset

The Result:

The Fast API web application has been developed with the Final model XGBoost with SHAP waterfall model to understand the most influential factors in Telecom customer churn.

http://localhost:8000/ - Link to go to the Webpage and End User (Telecom Industry) can select the options and "Predict" Button to check the result to below.

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

CHURN PREDICTION

City

Fresno

Senior Citizen

Yes

Partner

Yes

Dependents

No

Multiple Lines

No

Internet Service

DSL

Online Security

No

Online Backup

Yes

Device Protection

No

Tech Support

Yes

Streaming TV

No

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk

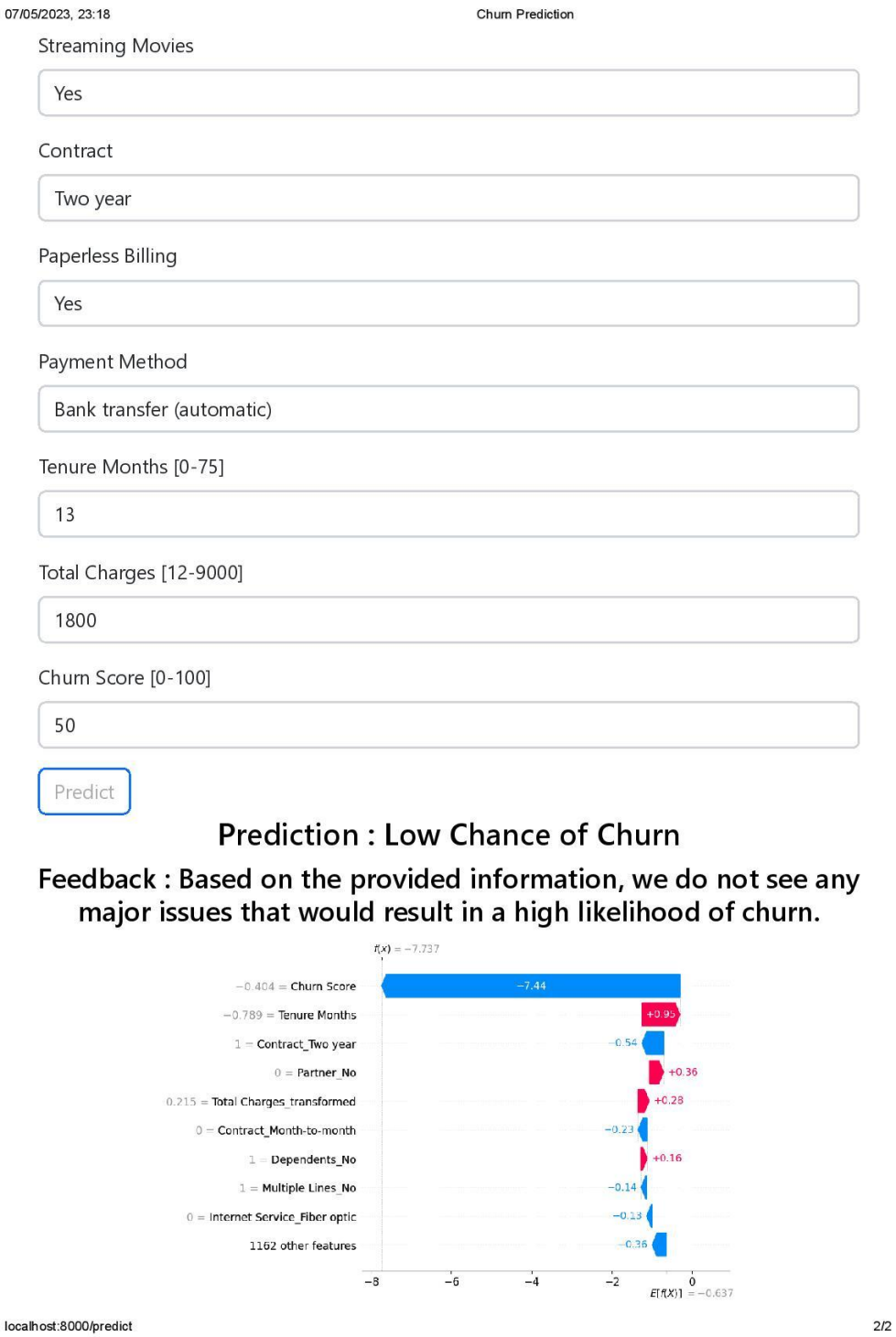


Fig: The FastAPI Web Application

Gowthami Nagappan - Teesside University, UK
B1326237@live.tees.ac.uk