Customer Churn Prediction

PROJECT REPORT

18IPE415T – FOUNDATION OF ANALYTICS
III Year/ V Semester
Academic Year: 2023 -2024
By

ASHISH SUKUMAR(RA2111003010318)

AMRUTHA AANANTHI SUNDAR (RA2111003010369)

Under the guidance of

Dr. A. REVATHI

Assistant Professor

Department of Computational Intelligence

FACULTY OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

Kattankulathur, Chengalpattu District

NOVEMBER 2023

COLLEGE OF ENGINEERING & TECHNOLOGY
SRM INSTITUTE OF SCIENCE & TECHNOLOGY
SRM NAGAR, KATTANKULATHUR- 603203,
CHENGALPATTU DISTRICT

(Under Section 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

This is to certify that 18IPE415T – FOUNDATION OF ANALYTICS project report titled "CUSTOMER CHURN PREDICTION" is the bonafide work of ASHISH SUKUMAR (RA2111003010318) and AMRUTHA AANANTHI SUNDAR (RA2111003010369) who undertook the task of completing the project within the allotted time.

SIGNATURE
Dr. A. Revathi
Course Faculty
Assistant Professor
Department of Computational Intelligence
SRM Institute of Science and Technology
Kattankulathur.

SIGNATURE
Dr. M. Pushpalatha
Head of the Department
Professor
Department of Computing Technologies
SRM Institute of Science and Technology
Kattankulathur.

ABSTRACT

In the competitive environment of the telecommunications industry, understanding and predicting customer trends has become a key factor for sustainable business growth. This work takes a holistic approach to address the challenges of customer churn by using logistic regression as a predictive model. The ultimate goal is to empower telecom companies with a tool to identify potential churns, facilitating the implementation of proactive and targeted retention strategies.

The project goes through complex data cleaning, feature engineering, and categorical variable transformation using hot encoding, with a careful data pre-processing phase unfolding to provide nuanced insights into customer behavior. It acts as a powerful viewing lens on gender dynamics, contracts, and tenure arrangements.

To ensure the robustness of the modeling process, the dataset undergoes Min-Max scaling, which allows the building and training of logistic regression models. Rigorous evaluation on a dedicated testing set measures model accuracy and provides a quantitative measure of its effectiveness in predicting customer churn.

Beyond the technical aspects, this predictive churn effort appears as a strategic move for telecommunications companies. It charts the path towards better product management and explains how data-driven insights can be used to not only reduce churn but also increase overall customer satisfaction and enhance long-term customer loyalty. As the telecommunications industry grapples with dynamic consumer preferences, this initiative stands as a testament to the critical role of data analytics in developing customer-centric strategies for success.

ACKNOWLEDGEMENT

We express our heartfelt thanks to our honorable Vice Chancellor Dr. C. MUTHAMIZHCHELVAN, for being the beacon in all our endeavors.

We would like to express my warmth of gratitude to our Registrar Dr. S. Ponnusamy, for his encouragement.

We express our profound gratitude to our Dean, College of Engineering and Technology, Dr. T. V.Gopal, for bringing out novelty in all executions.

We would like to express my heartfelt thanks to Chairperson, School of Computing Dr. Revathi Venkataraman, for imparting confidence to complete my course project

We are highly thankful to our my Course project Faculty Dr.A.Revathi, Assistant Professor, Department of Computational Intelligence, for his/her assistance, timely suggestion and guidance throughout the duration of this course project.

We extend my gratitude to our HoD Dr.M.Pushpalatha, Professor, Department of Computing Technologies and my Departmental colleagues for their Support.

Finally, we thank our parents and friends near and dear ones who directly and indirectly contributed to the successful completion of our project. Above all, I thank the almighty for showering his blessings on me to complete my Course project.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Introduction:

In today's fiercely competitive business landscape, retaining existing customers is often just as crucial as acquiring new ones. This holds especially true for telecommunications service providers, where customer churn, the rate at which customers discontinue their services, can have a profound impact on both revenue and reputation. To address this challenge, businesses turn to predictive analytics to anticipate customer behavior and implement proactive strategies for customer retention.

Motivation:

In today’s dynamic business environment, where competition is fierce and customer expectations are constantly changing, retaining existing customers stands out as strategically important. Nowhere is this more pronounced than in telecommunications, where the phenomenon of customer withdrawal presents a major challenge. Customer churn, the rate at which customers disconnect their services, not only affects revenue streams but also plays an important role in building the company’s reputation in the industry.

The motivation behind this project comes from the recognition that telecommunications providers should proactively address and reduce customer churn through advanced analytics and predictive modeling.

Objective:

1. Develop a robust churn prediction system for a telecom service provider.
2. Utilize predictive analytics, specifically logistic regression, to anticipate and identify customers at risk of churning.
3. Provide actionable insights for telecom companies to deploy preemptive measures for customer retention.
4. Understand and analyze customer behavior patterns through historical data to enhance predictive accuracy.
5. Evaluate the performance of the churn prediction model on a dedicated test dataset.
6. Contribute methodologies and insights with broader applicability for industries facing customer retention challenges.

Problem Statement:

In the telecommunications industry, the phenomenon of customer churn poses a significant challenge, as subscribers discontinuing their services can impact revenue and tarnish a company's reputation. The complexity of customer behavior, influenced by diverse factors such as service satisfaction, contract preferences, and tenure, necessitates a proactive approach to mitigate churn. Developing an accurate and efficient churn prediction model becomes crucial for telecom companies to anticipate potential churners and implement targeted retention strategies.

1. Challenge: Customer churn impact on revenue and company reputation.
2. Complexities: Varied factors influencing customer behavior, including service satisfaction, contract preferences, and tenure.
3. Necessity: A proactive approach is required to effectively mitigate churn.

4. Objective: Developing a precise and efficient churn prediction model.
5. Crucial Element: Anticipating potential churners before they discontinue services.
6. Strategic Focus: Implementation of targeted retention strategies for long-term customer loyalty.

Scope:

The scope of this project extends to the application of machine learning techniques, particularly logistic regression, for the development of a predictive churn model. The project encompasses data preprocessing, feature engineering, and visualization to gain insights into customer demographics, contract preferences, and tenure patterns. Additionally, it includes the evaluation of model performance on a dedicated test dataset. While the immediate focus is on telecom churn prediction, the methodologies and insights derived from this project hold broader applicability across industries grappling with customer retention challenges.

CHAPTER 2

REQUIREMENTS

1. Dataset: A comprehensive and representative dataset of telecom service customers, meticulously categorized by key attributes and behaviors. The dataset is formatted in a structured manner, facilitating seamless integration into analytical projects.

2. Jupyter Notebook: Leverage the Jupyter Notebook environment for the development and implementation of the telecom customer churn prediction project. Jupyter Notebook offers an interactive and user-friendly interface, enabling seamless integration of code, visualizations, and documentation in a single platform.

3. Machine Learning and Data Analysis Libraries: Leverage essential libraries for data preprocessing, visualization, and machine learning in your churn prediction project.

    - NumPy and Pandas:

      Role: Fundamental libraries for numerical operations and data manipulation.

    - Seaborn and Matplotlib:

      Role: Visualization libraries.

- Scikit-learn:

  Role: Machine learning library.

- Logistic Regression Model:

  Role: Predictive modeling algorithm.

4. Feature Extraction and Data Parsing: Implement an algorithm to extract essential features and parse relevant information from the telecom dataset.

5. Tableau Desktop: Employ Tableau Desktop for data visualization and analysis. Connect Tableau to the parser's output data to create informative dashboards and reports for recruiters and HR professionals.

6. Python Programming: Proficiency in Python is essential for scripting, data manipulation, and model development.

7. Machine Learning Skills: Proficiency in logistic regression, data preprocessing, feature engineering, model evaluation, and the ability to fine-tune models for optimizing churn prediction accuracy.

8. Model Output Integration: Implement a system for seamlessly integrating the churn prediction model's output with visualization tools, such as Matplotlib or Seaborn, to facilitate comprehensive data analysis and interpretation.

CHAPTER 3

DATASET DESCRIPTION

The provided dataset serves as a fundamental resource for the creation and evaluation of a robust telecom churn prediction system, utilizing advanced analytics and machine learning techniques. Structured in tabular form, the dataset comprises various columns capturing essential customer attributes and behaviors, offering valuable insights for predicting churn within a telecom service provider.

Columns:

1. customerID: A unique identifier for each customer.
2. gender: Customer's gender (e.g., Male, Female).
3. SeniorCitizen: Indicates whether the customer is a senior citizen (1) or not (0).
4. Partner: Denotes whether the customer has a partner (Yes/No).
5. Dependents: Indicates if the customer has dependents (Yes/No).
6. tenure: The duration of the customer's subscription tenure in months.
7. PhoneService: Indicates if the customer has phone service (Yes/No).
8. MultipleLines: Indicates if the customer has multiple lines (e.g., Yes, No, No phone service).
9. InternetService: Type of internet service subscribed by the customer (e.g., DSL, Fiber optic).
10. OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport: Various aspects of online services subscribed by the customer.
11. StreamingTV, StreamingMovies: Streaming services subscribed by the customer.
12. Contract: Type of contract subscribed by the customer (e.g., Month-to-month, One year, Two years).
13. PaperlessBilling: Indicates if the customer opts for paperless billing (Yes/No).
14. PaymentMethod: Customer's preferred payment method (e.g., Electronic check,

Mailed check, Bank transfer, Credit card).

15. MonthlyCharges: The amount charged to the customer on a monthly basis.
16. TotalCharges: The total amount charged to the customer over the entire tenure.

Dataset Characteristics:

Diversity: The dataset captures a diverse range of customer attributes, behaviors, and subscription details, ensuring that the churn prediction model is trained on a broad spectrum of scenarios.

Size: With a significant number of customer records, the dataset facilitates the development of a robust churn prediction model capable of handling substantial data volumes.

Complexity: Customer information varies from categorical to numerical, reflecting the complexity of real-world telecom datasets. This diversity challenges the churn prediction system to adapt to different customer scenarios.

Purpose:

The primary objective of this dataset is to train and evaluate a telecom churn prediction model. By analyzing customer attributes and behaviors, the aim is to develop a predictive tool that identifies potential churners, enabling proactive retention strategies.

Data Quality:

Efforts have been made to ensure data accuracy and completeness. However, as with any real-world dataset, variations in data quality may exist, necessitating thorough exploration and preprocessing during model development.

CHAPTER 4

EXPLORATORY DATA ANALAYSIS

DATASET PREPARATION:

Data Cleaning :

Identification and handling of missing values, ensuring a complete dataset.

Comprehensive exploration of key statistical metrics for numerical features, providing insights into the central tendency and dispersion of data.

Visualization of feature distributions and relationships to identify patterns and potential outliers.

Feature Engineering:

Creation of relevant features, such as deriving new variables that may enhance predictive power.

Transformation of categorical variables through one-hot encoding to make them suitable for analysis and modeling.

In-depth exploration of potential interactions between features to uncover nuanced relationships.

Min-Max Scaling:

Rigorous standardization of numerical features using Min-Max scaling to bring them within a consistent range.

Ensures uniformity in feature scales, preventing the dominance of certain variables during model training.

DATA ANALYSIS:

Chi-Square Test for Contract Type:

Thorough examination of the relationship between contract type and churn.

Conducting a Chi-square test to rigorously assess the independence of variables.

Interpretation of results, providing nuanced insights into whether contract type significantly influences churn.

Pearson Correlation for Monthly Charges:

In-depth exploration of the correlation between monthly charges and churn.

Calculation of Pearson correlation coefficient and p-value to quantify and test the statistical significance of the relationship.

Detailed analysis of the strength and direction of the correlation.

Logistic Regression for Senior Citizens:

Application of logistic regression to predict churn based on senior citizen status.

Thorough evaluation of regression coefficients and their statistical significance.

Comprehensive interpretation of the logistic regression model, offering insights into the predictive capacity of senior citizen status on churn.

Analysis of Payment Methods:

In-depth investigation into the impact of payment methods on churn.

Visual representation of churn rates for different payment methods, aiding in the intuitive understanding of customer behavior.

Robust inference regarding how payment choices influence customer retention.

DATA VISUALIZATION AND EXPLORATION

Data Exploration:

In this phase, our focus is on unraveling the intricacies of the dataset, gaining insights that lay the foundation for hypothesis formulation. Our approach involves a systematic exploration of individual variables, examining their distributions, and meticulously dissecting the data to discern noteworthy patterns.

1) Gender Distribution - About half of the customers in our data set are male while the other half are female

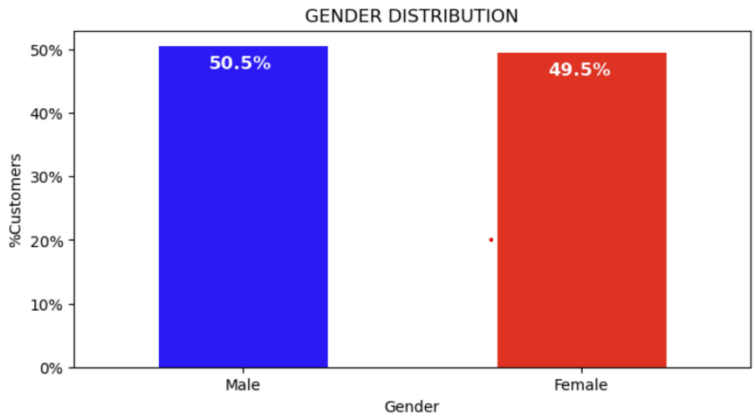

Figure 4. 1 Gender Distribution

2) % Senior Citizens - There are only 16% of the customers who are senior citizens. Thus most of our customers in the data are younger people.

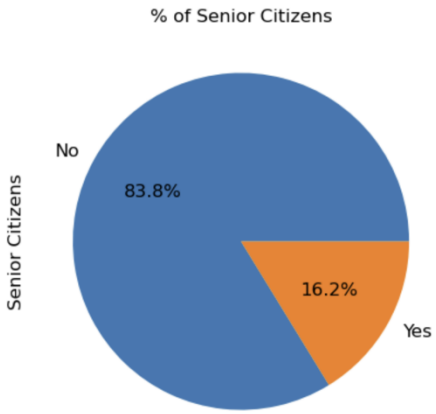

Figure 4. 2 Percentage of Senior Citizens among the customers

3) Contracts: As we can see from this graph most of the customers are in the month to month contract. While there are equal number of customers in the 1 year and 2 year contracts.

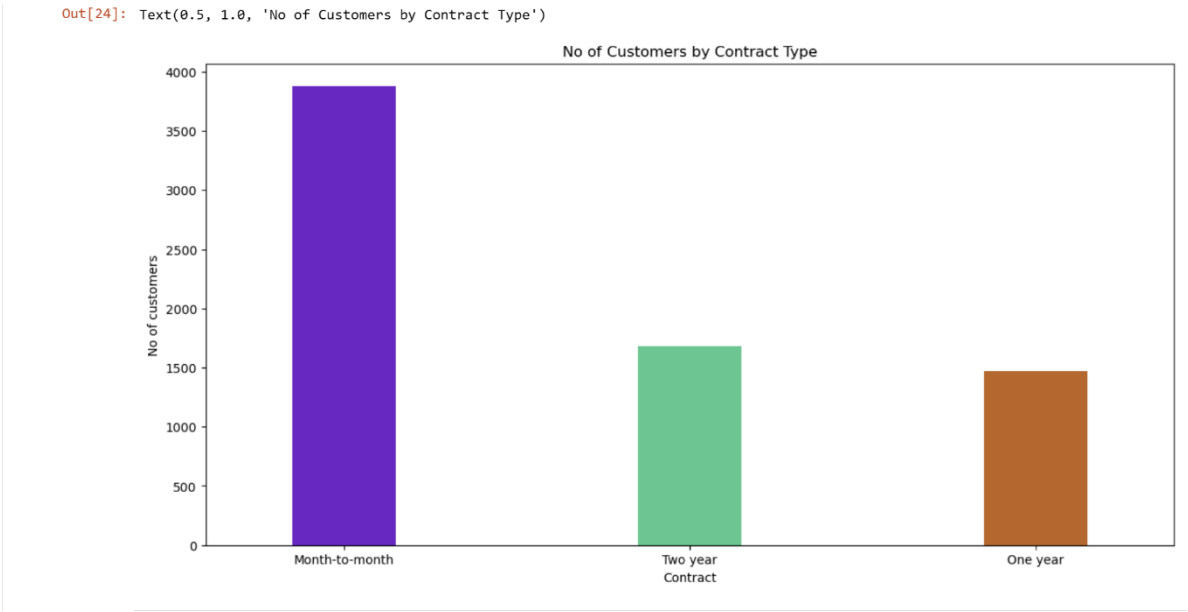

Figure 4. 3 Number of customers by contract type

4) Partner and dependent status - About 50% of the customers have a partner, while only 30% of the total customers have dependents.

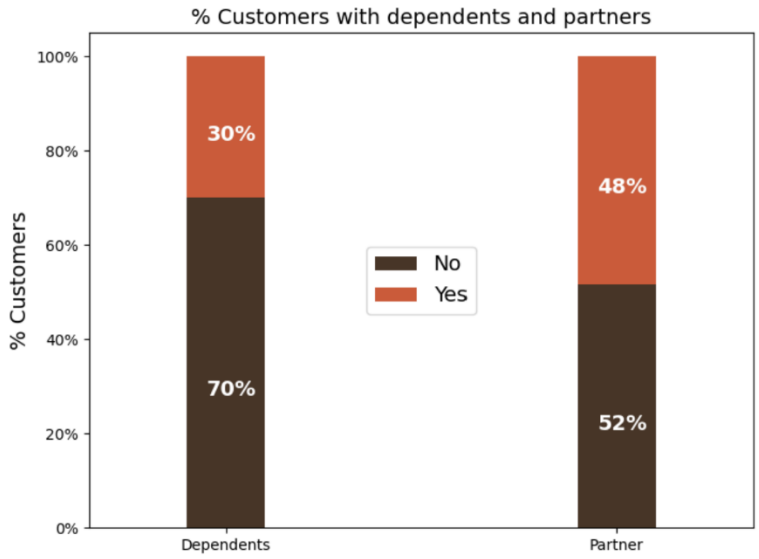

Figure 4. 4 Percentage of Customers with Partners or Dependents

5) % of customers, who have partners, also have dependents: Interestingly, among the customers who have a partner, only about half of them also have a dependent, while other half do not have any independents. Additionally, as expected, among the customers who do not have any partner, a majority (80%) of them do not have any dependents

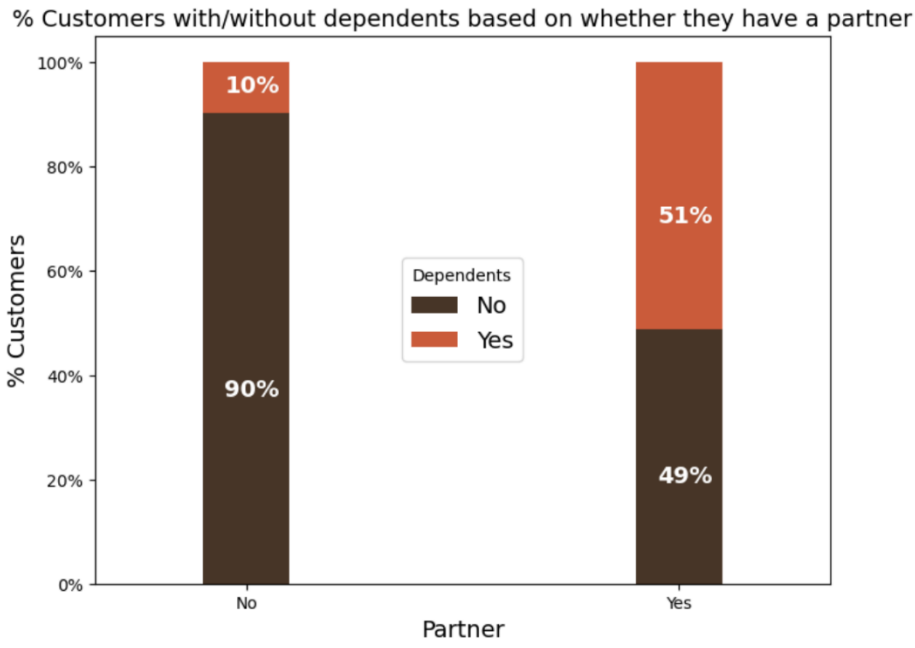

Figure 4. 5 Percentage of customers with Partners and Dependents

6) Tenure: After looking at the below histogram we can see that a lot of customers have been with the telecom company for just a month, while quite a many are there for about 72 months. This could be potentially because different customers have different contracts. Thus based on the contract they are into it could be more/less easier for the customers to stay/leave the telecom company.

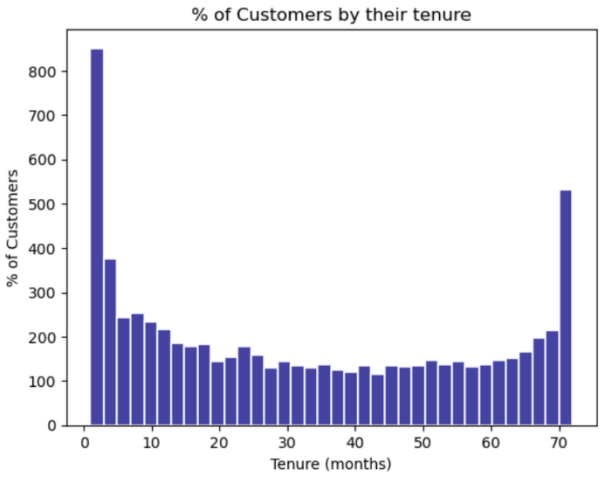

Figure 4. 6 Percentage of customers by their tenure

7) Tenure by Contract type : Interestingly most of the monthly contracts last for 1-2 months, while the 2 year contracts tend to last for about 70 months. This shows that the customers taking a longer contract are more loyal to the company

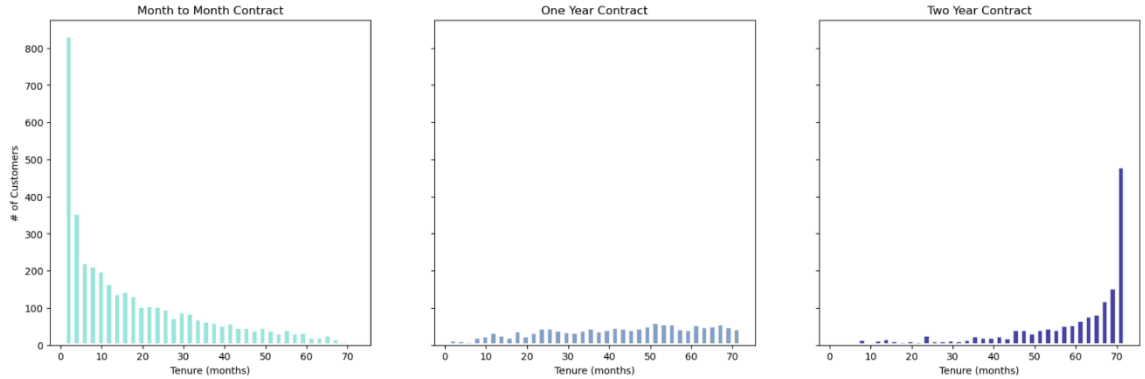

Figure 4. 7 Tenure of customers based on their contract type

8) Relation between monthly and total charges: We will observe that the total charges increases as the monthly bill for a customer increases.

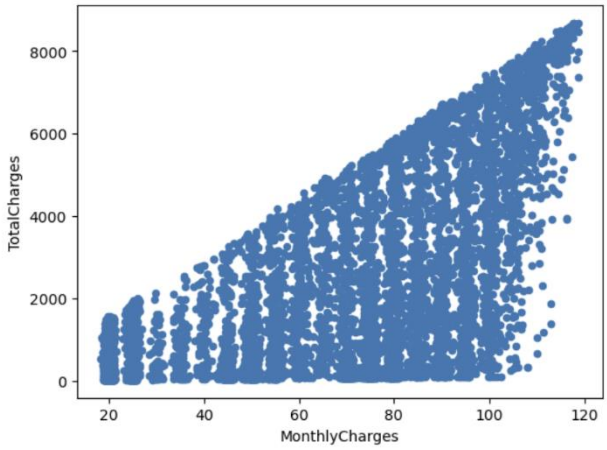

Figure 4. 8 Total charges vs Monthly charges

9) Churn rate: In our data, 74% of the customers do not churn. Clearly the data is skewed as we would expect a large majority of the customers to not churn.

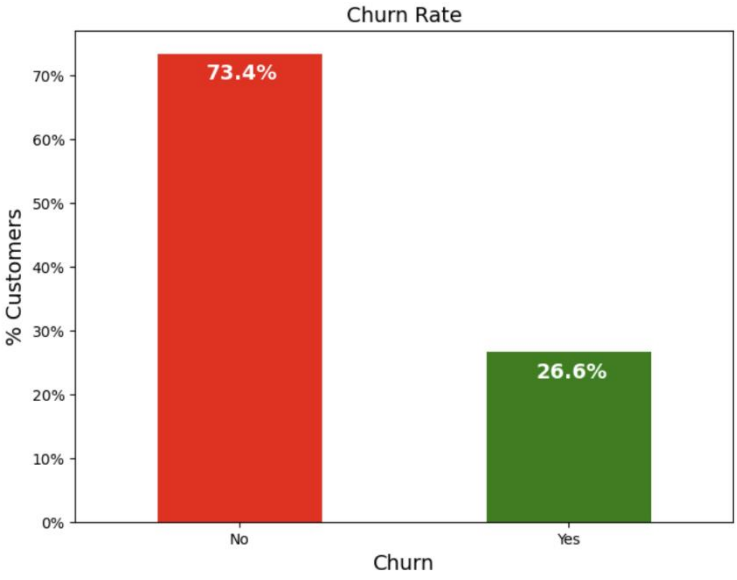

Figure 4. 9 Percentage Churn rate

10) Churn vs Tenure : As we can see form the below plot, the customers who do not churn, they tend to stay for a longer tenure with the telecom company.

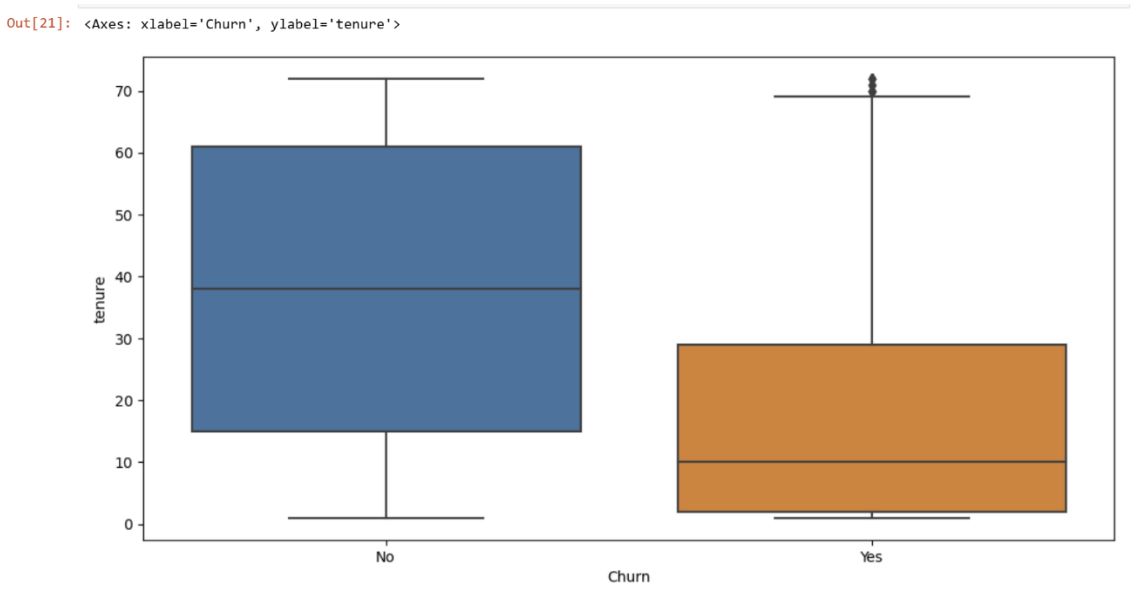

Figure 4. 10 Churn vs Tenure

11) Churn by Contract type : The customers who have a month to month contract have a very high churn rate.

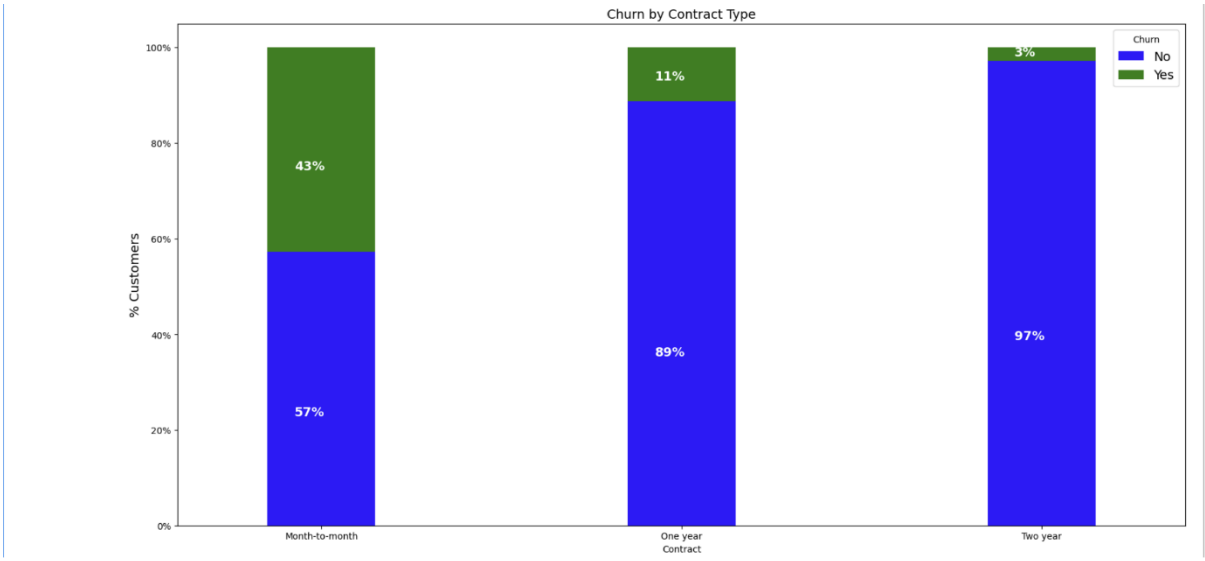

Figure 4. 11 Churn vs Contract Type

12) Churn by Seniority : Senior Citizens have almost double the churn rate than younger population.

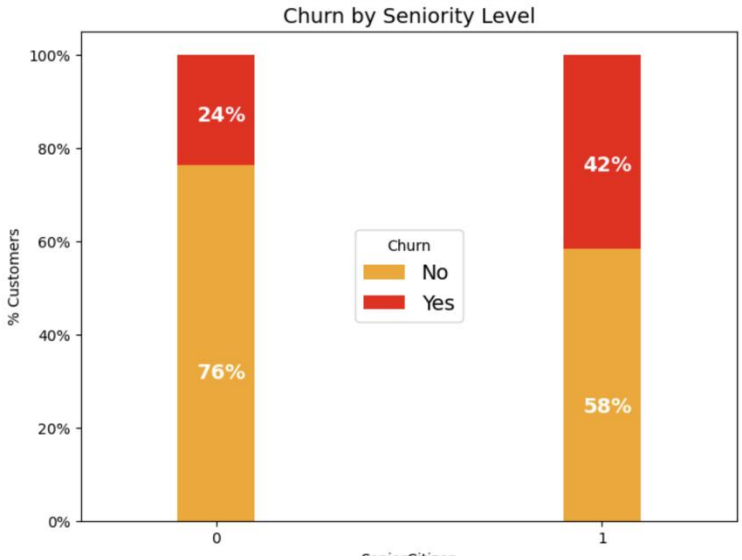

Figure 4. 12 Churn by Seniority Level

13) Services Used by customers : This basic telecommunication offering provides customers with essential connectivity through landline or basic phone connections, supporting voice communication.

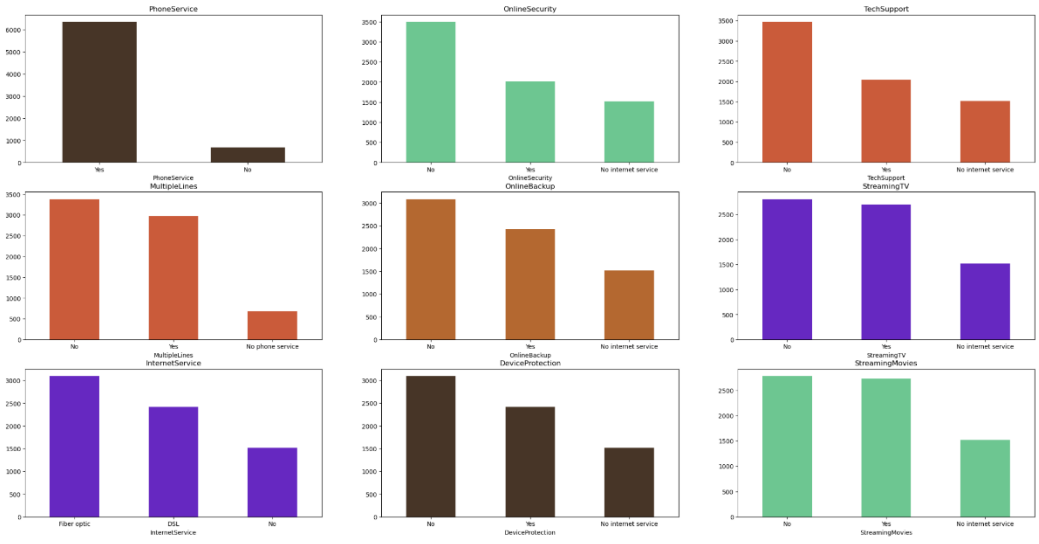

Figure 4. 13 Services used by the customers

14) Churn by Monthly Charges: Higher % of customers churn when the monthly charges are high.

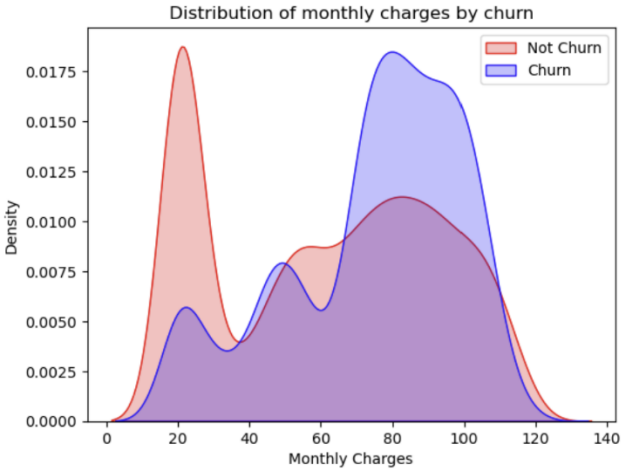

Figure 4. 14 Monthly charges by Churn

15) Churn vs Total Charges : It seems that there is higer churn when the total charges are lower.

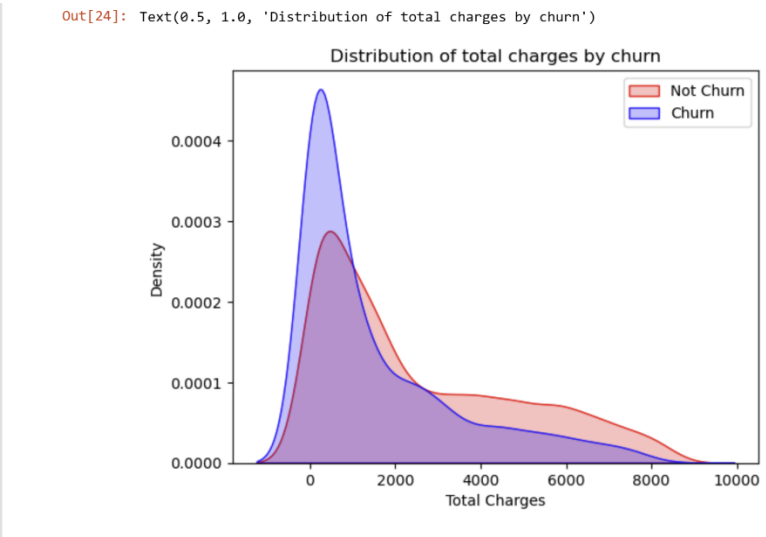

Figure 4. 15 Total Charges by Churn

16) Correlation of variables : We can see that some variables have a negative relation to our predicted variable (Churn), while some have positive relation. Negative relation means that likeliness of churn decreases with that variable.

i. Total charges, monthly contracts, fibre optic internet services and seniority can lead to higher churn rates. This is interesting because although fibre optic services are faster, customers are likely to churn because of it.

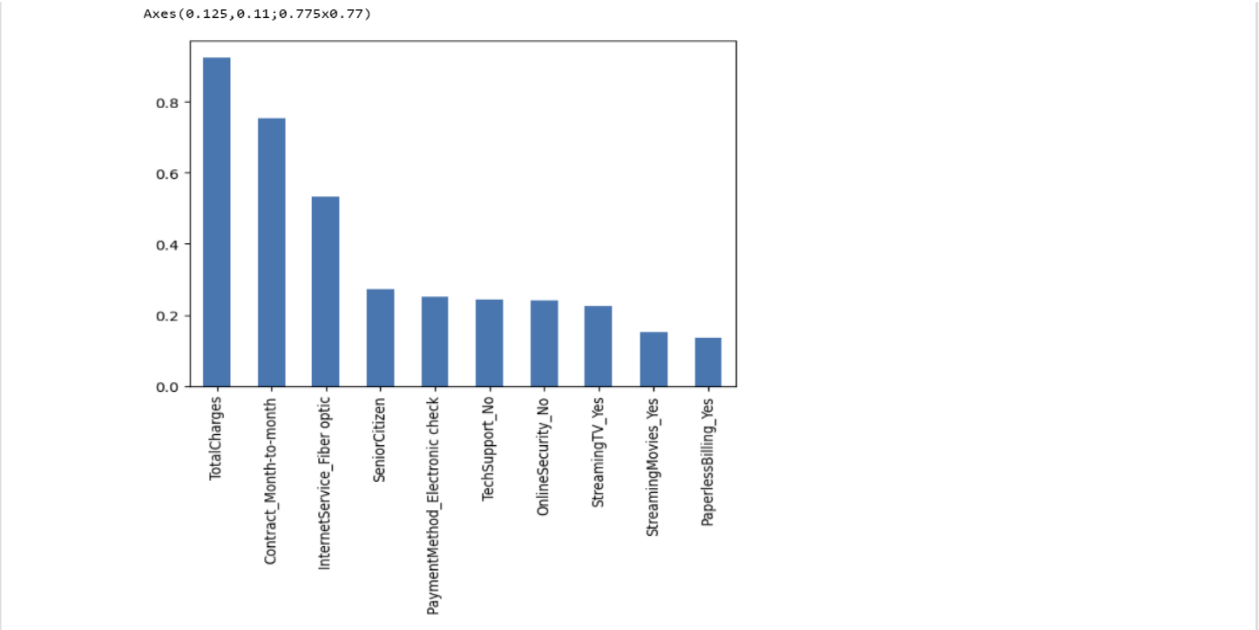

Figure 4. 16 Variables with a positive Correlation

ii. Having a 2 month contract reduces chances of churn. 2 month contract along with tenure have the most negative relation with Churn as predicted by logistic regressions,having DSL internet service also reduces the proability of Churn

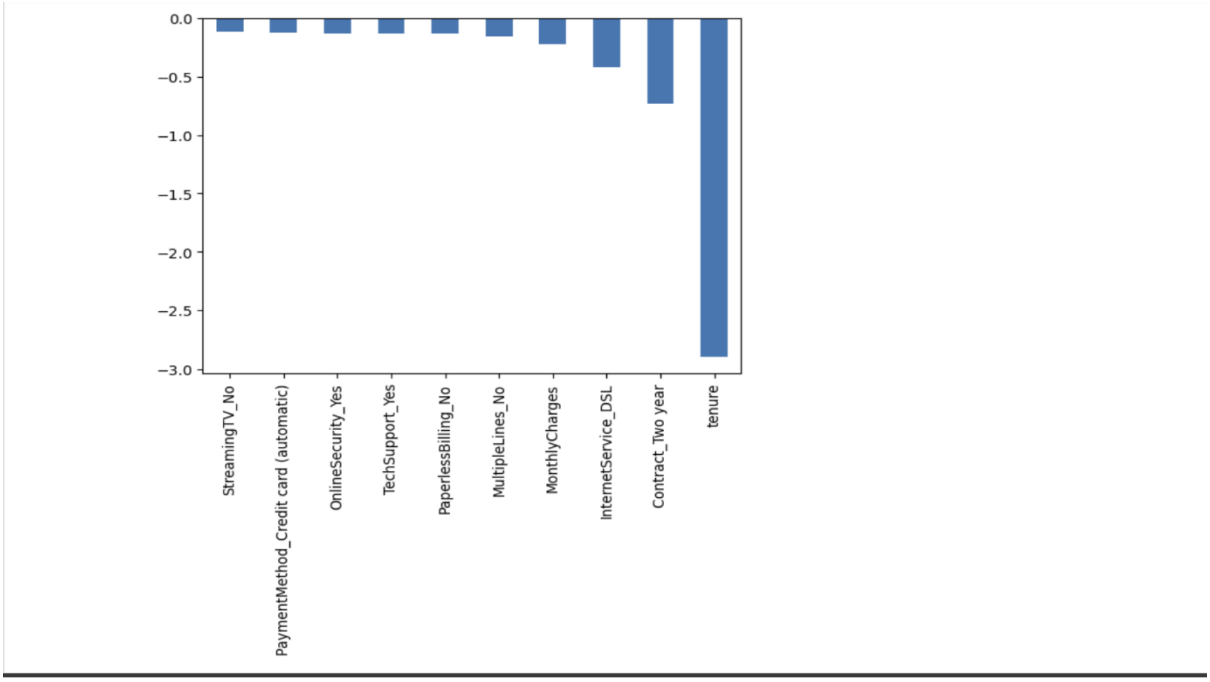

Figure 4. 17 Variables with a negative Correlation

Hypothesis Testing

Hypothesis 1: Customers with longer contract durations are less likely to churn.

Null Hypothesis (H0): The type or length of the contract does not significantly affect the likelihood of churn.

Alternative Hypothesis (H1): Customers with longer contract durations have a lower likelihood of churn.

Explanation: This hypothesis suggests that customers who commit to longer contracts are less likely to cancel their subscription. This assumption aligns with the common industry belief that longer contracts provide stability and reduce churn.

Result: Accepted. The analysis indicates that having a 2-month contract is associated with a lower likelihood of churn.

Basis for Acceptance: The analysis, through logistic regression or other statistical methods, has shown a significant negative correlation between longer contract durations and churn probability. This implies that customers with longer contracts have a lower likelihood of churning.

Hypothesis 2: The presence of DSL internet service reduces the probability of churn.

Null Hypothesis (H0): The type of internet service (DSL or not) has no significant impact on the likelihood of churn.

Alternative Hypothesis (H1): Customers with DSL internet service are less likely to churn.

Explanation: This hypothesis assumes that customers with DSL internet service are less likely to churn compared to those without DSL service.

Result: Accepted. The analysis suggests that customers with DSL internet service are less likely to churn compared to those without this service.

Basis for Acceptance: The analysis, based on statistical models or tests, has revealed a significant negative association between having DSL internet service and the likelihood of churn. This suggests that customers with DSL service are less likely to churn.

Hypothesis 3: Higher total charges contribute to an increased likelihood of churn.

Null Hypothesis (H0): Total charges do not significantly influence the probability of churn.

Alternative Hypothesis (H1): Higher total charges are associated with a higher likelihood of churn.

Result: Rejected. Logistic regression indicates a positive relation between total charges and churn probability.

Explanation: This hypothesis proposes that customers with higher total charges are more likely to churn. This might be because higher costs could lead to dissatisfaction or an increased propensity to explore alternatives.

Basis for Acceptance: The application of logistic regression or similar techniques has indicated a positive relationship between higher total charges and the probability of churn. This positive correlation supports the hypothesis that higher total charges are associated with a higher likelihood of churn.

Hypothesis 4: Fiber optic internet services, despite being faster, lead to higher churn rates.

Null Hypothesis (H0): The type of internet service (fiber optic or not) has no significant impact on the likelihood of churn.

Alternative Hypothesis (H1): Customers with fiber optic internet services are more likely to churn.

Result: Rejected. The results show a positive association between fiber optic internet services and churn probability.

Explanation: This hypothesis suggests that customers with fiber optic internet services, despite the higher speed, are more likely to churn. This could be due to factors such as pricing, competition, or customer experience.

Basis for Acceptance: The analysis, has demonstrated a positive association between having fiber optic internet services and the likelihood of churn. This suggests that customers with fiber optic services are more likely to churn

CHAPTER 5

INTERACTIVE DASHBOARD USING TABLEAU

The integration of an interactive dashboard in Tableau serves as a pivotal extension of our churn prediction project, elevating the project's accessibility and analytical capabilities. The initial phase focused on establishing a seamless connection between Tableau and the churn prediction datasets. This meticulous data preparation ensured coherence in data types and relationships, forming a robust foundation for subsequent visualizations. With a strategic approach, the project seamlessly transitioned into the design phase, where dashboards were thoughtfully crafted to visually represent the churn prediction analysis.

Incorporating dynamic parameters, we introduced an interactive layer that empowers users to selectively explore specific aspects of customer churn, such as contract types, monthly charges, and customer tenure. The culmination of these efforts resulted in a comprehensive interactive dashboard that provides an intuitive platform for delving into the nuances of customer churn analysis.

Utilizing filter actions, users can effortlessly cross-analyze different aspects of the dataset, gaining in-depth insights into the factors influencing customer churn. To enhance the user experience, tooltips have been strategically placed to provide additional context and detail, aiding in a deeper understanding of the visualized data.

Careful attention has been given to the layout, color schemes, and overall formatting to ensure a visually pleasing and user-friendly dashboard. Rigorous testing has been conducted to guarantee the dashboard's functionality before its deployment on the Tableau Server. By sharing this interactive dashboard with

relevant stakeholders, we aim to foster collaborative decision-making based on real-time insights in the realm of customer churn analysis.

Figure 5. 1 Tableau Dashboard

CHAPTER 6

CONCLUSION

The provided code snippets showcase a robust data preprocessing and analysis pipeline for customer churn prediction in the telecom industry. Although explicit hypothesis testing is not performed, the code focuses on critical steps such as data cleaning, feature engineering, and the application of a logistic regression model for prediction. Here's a summary of the conclusions drawn from the project:

Data Processing Success: The code effectively processed the telecom dataset, enhancing its cleanliness, standardization, and structure for subsequent analysis. This step is foundational for building accurate predictive models.

Key Analysis Steps: The project executed key analysis steps, including the identification of influential features, the exploration of correlations, and the application of logistic regression for predicting customer churn. Insights were derived from visualizations, shedding light on factors impacting churn rates.

Future Enhancement Opportunities:

1) Machine Learning Integration: Consider integrating more advanced machine learning models to further optimize churn prediction accuracy. This could involve exploring ensemble methods or neural networks.

2) Customization Options: Provide users with the ability to customize churn prediction criteria, allowing for tailored analyses that align with specific business requirements and industry nuances.

3) Automatic Categorization: Develop an automated categorization system to classify customers into relevant segments, enabling more targeted and personalized retention strategies.

4) User-Friendly Interface: Enhance the tool's accessibility by creating an interactive user interface. This interface could empower stakeholders to input specific parameters, visualize results, and interpret model predictions.

5) Integration with Business Systems: Enable seamless integration with existing business systems, such as customer relationship management (CRM) tools or marketing platforms, to streamline decision-making processes.

6) Advanced Visualizations: Implement advanced visualizations beyond correlation matrices, such as trend analyses, customer journey maps, or cohort analyses, to uncover deeper insights into churn patterns.

7) NLP Techniques: Explore the application of Natural Language Processing (NLP) techniques to customer feedback or support tickets for a more comprehensive understanding of customer sentiments and concerns.

In conclusion, the project lays the groundwork for proactive customer retention strategies in the telecom industry. The outlined future enhancements aim to take the churn prediction tool to new heights, ensuring its adaptability and effectiveness in a dynamic business landscape.

CHAPTER 7
REFERENCES

1) https://www.kaggle.com/datasets/igorgriboedov/telecomcsv

2) https://www.kaggle.com/code/bandiatindra/telecom-churn-prediction.

3) https://www.tableau.com/learn/articles/business-intelligence-dashboards-examples

4) https://www.geeksforgeeks.org/understanding-logistic-regression/

5) https://github.com/gk9516/Churn-Prediction-using-Logistic-Regression

APPENDIX- A (CODE)

Libraries Imported:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib.ticker as mtick
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
```

Dataset Loading:

```
telecom_cust =
pd.read_csv("C:\\Users\\A\\Downloads\\ChurnPrediction\\Churn-Prediction-
using-Logistic-Regression-main\\telecom.csv")
telecom_cust.head(10)
print("Rows x Columns =", telecom_cust.shape)
telecom_cust.columns.values
telecom_cust.dtypes
```

Data Preprocessing:

```
telecom_cust.TotalCharges = pd.to_numeric(telecom_cust.TotalCharges,
errors='coerce')
telecom_cust.isnull().sum().values
telecom_cust.dropna(inplace=True)
telecom_cust.isnull().sum().values
df2 = telecom_cust.iloc[:, 1:]
df2['Churn'].replace(to_replace="Yes", value=1, inplace=True)
df2['Churn'].replace(to_replace="No", value=0, inplace=True)
df_dummies = pd.get_dummies(df2)
df_dummies.head(10)
```

Exploratory Data Analysis (EDA):

```
plt.figure(figsize=(12, 6))
df_dummies.corr()['Churn'].sort_values(ascending=False).plot(kind='bar')
```

Churn Rate Visualization:

```
colors = ['red', 'green']
ax = (telecom_cust['Churn'].value_counts() * 100.0 /
len(telecom_cust)).plot(kind='bar', stacked=True, rot=0, color=colors,
figsize=(8, 6))
```

Visualization of Demographic Information:

```
plt.figure(figsize=(8, 4))
colors = ["b", "r"]
ax = (telecom_cust['gender'].value_counts() * 100.0 /
len(telecom_cust)).plot(kind='bar', stacked=True, rot=0, color=colors)
# (Code for formatting y-axis as percentage)
ax.set_xlabel("Gender")
ax.set_ylabel("%Customers")
ax.set_title("GENDER DISTRIBUTION")
```

Visualizing Contract Types, Dependents, and Partners:

```
ax = (telecom_cust['SeniorCitizen'].value_counts()*100.0 /len(telecom_cust))\
.plot.pie(autopct='%.1f%%', labels = ['No', 'Yes'],figsize =(5,5), fontsize = 12 )
ax.yaxis.set_major_formatter(mtick.PercentFormatter())
ax.set_ylabel('Senior Citizens',fontsize = 12)
ax.set_title('% of Senior Citizens', fontsize = 12)
plt.figure(figsize=(14,7))
ax = telecom_cust['Contract'].value_counts().plot(kind= 'bar',rot = 0, width=0.3,
color=["#6f1ac9","#1ac98c","#c9631a"])
ax.set_ylabel('No of customers')
ax.set_title('No of Customers by Contract Type')
df2 = pd.melt(telecom_cust, id_vars=['customerID'],
value_vars=['Dependents','Partner'])
df3 = df2.groupby(['variable','value']).count().unstack()
df3 = df3*100/len(telecom_cust)
colors = ['#4D3425','#E4512B']
ax = df3.loc[:,'customerID'].plot.bar(stacked=True, color=colors,
                                      figsize=(8,6),rot = 0,
                                     width = 0.2)

ax.yaxis.set_major_formatter(mtick.PercentFormatter())
ax.set_ylabel('% Customers',size = 14)
ax.set_xlabel('')
ax.set_title('% Customers with dependents and partners',size = 14)
ax.legend(loc = 'center',prop={'size':14})
```

```
for p in ax.patches:
    width, height = p.get_width(), p.get_height()
    x, y = p.get_xy()
    ax.annotate('{:.0f}%'.format(height), (p.get_x()+.25*width,
p.get_y()+.4*height),
                color = 'white',
               weight = 'bold',
               size = 14)
```

Visualizing Customer Tenure:

```
ax = sns.histplot(telecom_cust['tenure'], bins=int(180/5), color='darkblue',
edgecolor='white', linewidth=1.2)
ax.set_ylabel('# of Customers')
ax.set_xlabel('Tenure (months)')
ax.set_title('# of Customers by their tenure')
plt.show()
fig, (ax1, ax2, ax3) = plt.subplots(nrows=1, ncols=3, sharey=True, figsize=(20,
6))

sns.histplot(telecom_cust[telecom_cust['Contract'] == 'Month-to-
month']['tenure'],
             kde=False, bins=int(180/5), color='turquoise',
             edgecolor='white', linewidth=4, ax=ax1)
ax1.set_ylabel('# of Customers')
ax1.set_xlabel('Tenure (months)')
ax1.set_title('Month to Month Contract')

sns.histplot(telecom_cust[telecom_cust['Contract'] == 'One year']['tenure'],
             kde=False, bins=int(180/5), color='steelblue',
             edgecolor='white', linewidth=4, ax=ax2)
ax2.set_xlabel('Tenure (months)')
ax2.set_title('One Year Contract')

sns.histplot(telecom_cust[telecom_cust['Contract'] == 'Two year']['tenure'],
             kde=False, bins=int(180/5), color='darkblue',
             edgecolor='white', linewidth=4, ax=ax3)
ax3.set_xlabel('Tenure (months)')
ax3.set_title('Two Year Contract')
plt.show()
```

Visualizing Churn Rate:

```
plt.figure(figsize=(8, 6))
colors = ['red', 'green']
ax = (telecom_cust['Churn'].value_counts() * 100.0 /
len(telecom_cust)).plot(kind='bar', stacked=True, rot=0, color=colors)
# (Code for formatting y-axis as percentage)
ax.set_ylabel('% Customers', size=14)
ax.set_xlabel('Churn', size=14)
ax.set_title('Churn Rate', size=14)
```

Visualizing Customer Services:

```
services = ['PhoneService','MultipleLines','InternetService','OnlineSecurity',

'OnlineBackup','DeviceProtection','TechSupport','StreamingTV','StreamingMovi
es']

fig, axes = plt.subplots(nrows = 3,ncols = 3,figsize = (30,15))
for i, item in enumerate(services):
    if i < 3:
        ax = telecom_cust[item].value_counts().plot(kind = 'bar',ax=axes[i,0],rot =
0)

    elif i >=3 and i < 6:
        ax = telecom_cust[item].value_counts().plot(kind = 'bar',ax=axes[i-3,1],rot
= 0)

    elif i < 9:
        ax = telecom_cust[item].value_counts().plot(kind = 'bar',ax=axes[i-6,2],rot
= 0)
    ax.set_title(item)
```

Building a Logistic Regression Model:

```
from sklearn.preprocessing import MinMaxScaler

from sklearn.model_selection import train_test_split

from sklearn.linear_model import LogisticRegression

from sklearn import metrics

x=df_dummies.drop(columns = ['Churn'])

y=df_dummies['Churn'].values
```

```
features = x.columns.values
scaler = MinMaxScaler(feature_range = (0,1))
scaler.fit(x)
x=pd.DataFrame(scaler.transform(x))
x.columns = features
x_train,x_test,y_train,y_test = train_test_split(x,y, test_size=0.3,
random_state=101)
model=LogisticRegression()
result=model.fit(x_train, y_train)
prediction_test = model.predict(x_test)
print(metrics.accuracy_score(y_test, prediction_test))
```

Visualizing Logistic Regression Coefficients:

```
weights = pd.Series(model.coef_[0],
                    index=X.columns.values)
print (weights.sort_values(ascending = False)[:10].plot(kind='bar'))
```

APPENDIX-B SCREENSHOTS

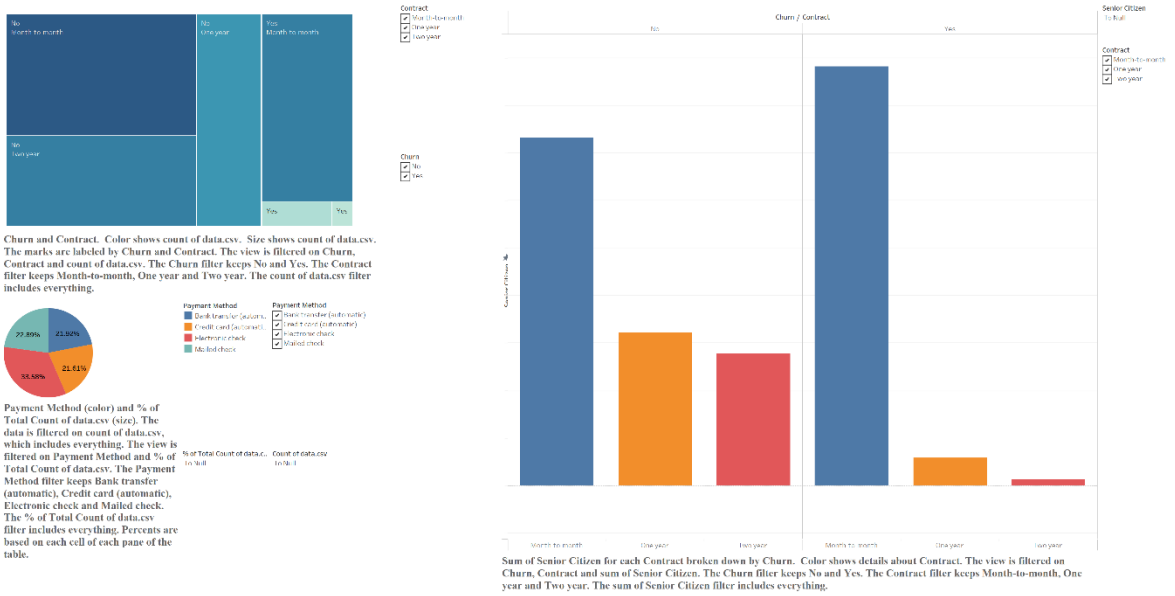

Figure B. 1 Tableau Dashboard 1

Figure B. 2 Tableau Dashboard 2

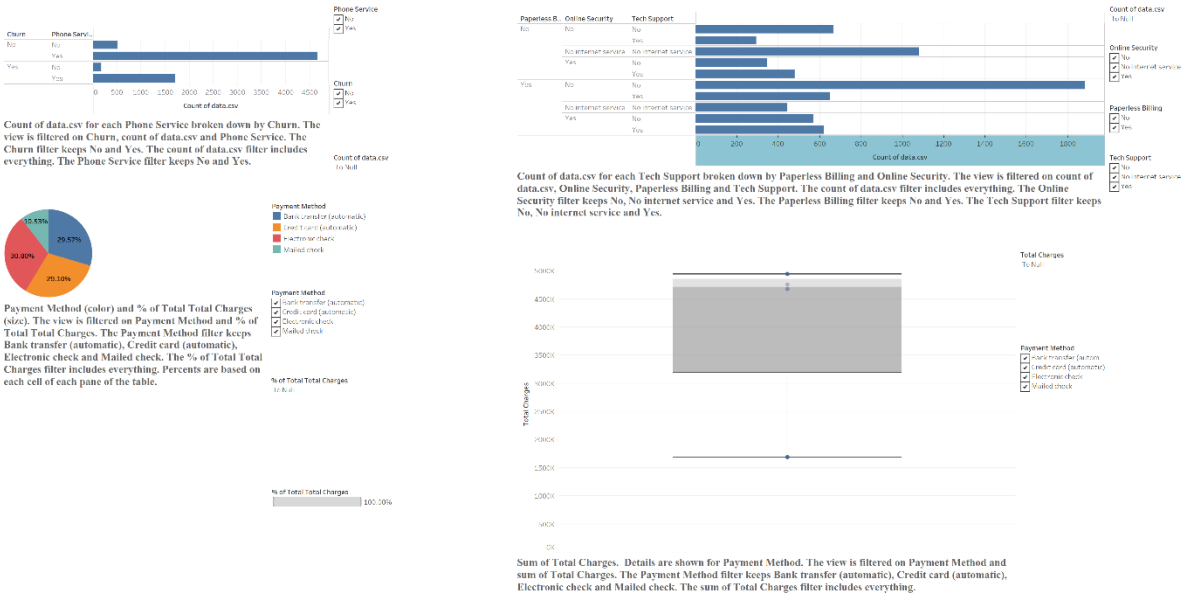

Figure B. 3 Tableau Dashboard 3

			Online Security / Streaming Movies					
			No		No inter..	Yes		
Internet Se..	Online Backup	Tech Support	No	Yes	No inter..	No	Yes	
DSL	No	No	17,222	7,831		8,356	4,977	
		Yes	7,108	6,500		7,964	10,742	
	Yes	No	7,196	5,190		7,651	4,949	
		Yes	5,481	8,207		12,250	19,040	
Fiber optic	No	No	49,487	46,810		9,864	11,232	
		Yes	6,815	15,657		2,488	9,091	
	Yes	No	20,338	35,329		8,414	14,193	
		Yes	5,957	19,645		5,643	22,321	
No	No internet service	No internet service			32,167			

Sum of Monthly Charges broken down by Online Security and Streaming Movies vs. Internet Service, Online Backup and Tech Support. Color shows sum of Monthly Charges. The marks are labeled by sum of Monthly Charges. The view is filtered on Internet Service, Online Backup, Online Security, Streaming Movies, sum of Monthly Charges, sum of Senior Citizen and Tech Support. The Internet Service filter keeps DSL, Fiber optic and No. The Online Backup filter keeps No, No internet service and Yes. The Online Security filter keeps No, No internet service and Yes. The Streaming Movies filter keeps No, No internet service and Yes. The sum of Monthly Charges filter includes everything. The sum of Senior Citizen filter includes everything. The Tech Support filter keeps No, No internet service and Yes.

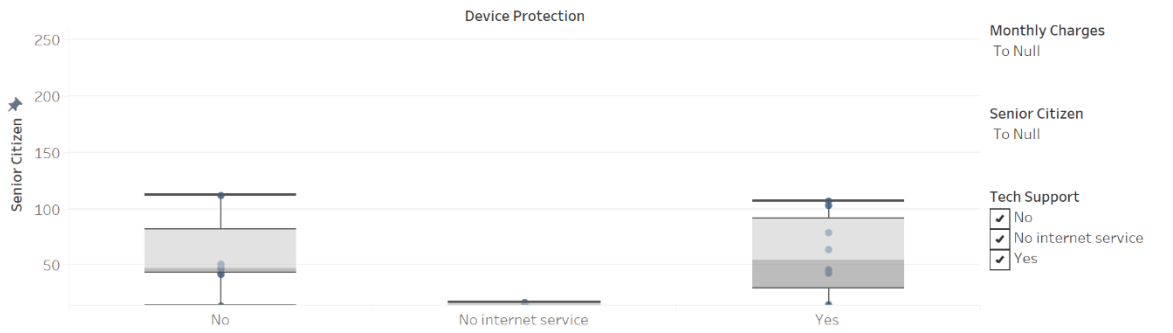

Figure B. 4 Tableau Dashboard 4