SUMMER TRAINING/INTERNSHIP

PROJECT REPORT

(Term June-July 2025)

TELECOMMUNICATION CUSTOMER CHURN ANALYSIS

Submitted by

Names: Anushka (12319243)
Prerna Agrawal (12325546)
Harpreet Singh (12316448)
Htet Aung Khant (12303558)
Gaurav Kumar (12317453)

Course Code : TETV76
Course Name : FROM DATA TO DECISIONS - A HANDS ON APPROACH TO DATA SCIENCE

Under the Guidance of

Sandeep Kaur (23614)

School of Computer Science and Engineering

ABSTRACT

This report presents a comprehensive analysis of customer churn prediction in the telecommunications industry using machine learning techniques and business intelligence tools. The project involves exploratory data analysis, feature engineering, implementation of multiple machine learning models including Support Vector Machines (SVM) and Logistic Regression, customer segmentation using K- Means clustering, and development of an interactive Power BI dashboard. The analysis aims to identify key factors contributing to customer churn and provide actionable insights for customer retention strategies. The results demonstrate the effectiveness of machine learning approaches in predicting customer churn with significant accuracy improvements through proper data preprocessing and model selection.

1.Introduction

Customer churn, defined as the rate at which customers discontinue their service subscriptions, represents a critical challenge for telecommunications companies. With the increasing competition in the telecommunications market, retaining existing customers has become more cost-effective than acquiring new ones. The ability to predict which customers are likely to churn enables companies to implement targeted retention strategies, thereby reducing revenue loss and improving customer lifetime value.

This project develops a comprehensive churn prediction system by analyzing customer behavior in telecom services, focusing on billing history, contract type, service usage, and support interactions. The methodology combines traditional statistical analysis with modern machine learning techniques to create predictive models that can identify at-risk customers with high accuracy.

2.Problem Statement

The primary objective of this project is to construct a churn prediction system by analyzing customer behavior in telecom services. The analysis examines multiple dimensions of customer data including billing history, contract type, service usage patterns, and support interactions. The project performs exploratory data analysis (EDA), data cleaning, machine learning model development, and creates an interactive dashboard for business stakeholders.

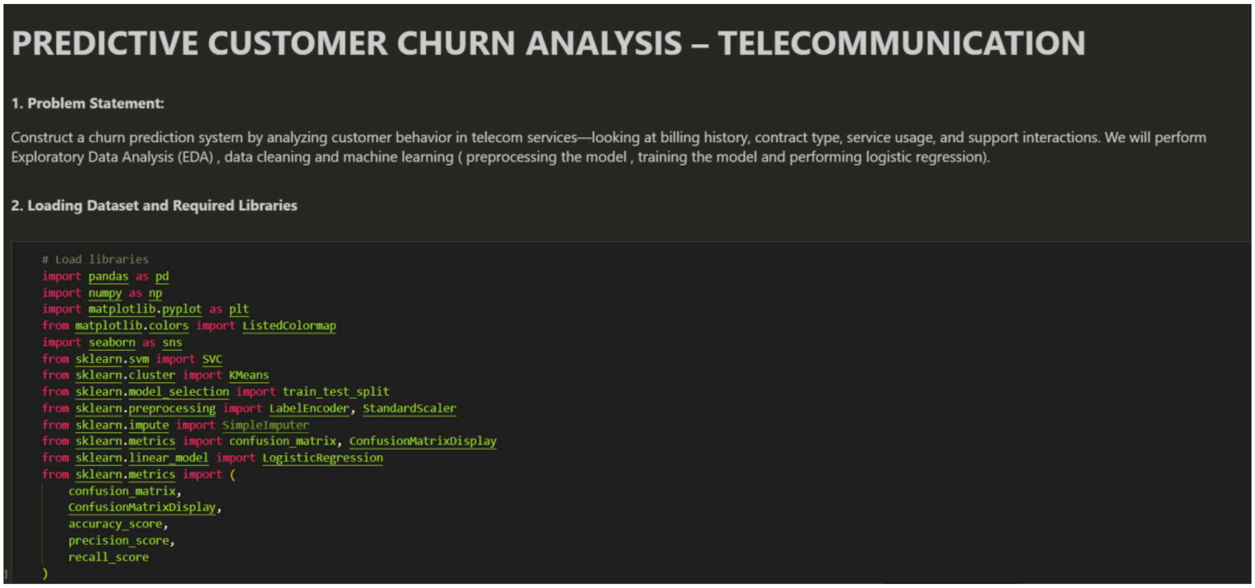
3.Methodology

3.1.Data Collection and Preprocessing

The analysis begins with comprehensive data preprocessing including handling missing values, feature scaling, and categorical variable encoding. The dataset contains customer demographic information, service usage patterns, billing data, and churn indicators.

3.2.Exploratory Data Analysis

Extensive exploratory data analysis was performed to understand the underlying patterns in the data. Key visualizations include distribution analysis, correlation studies, and customer segmentation analysis.



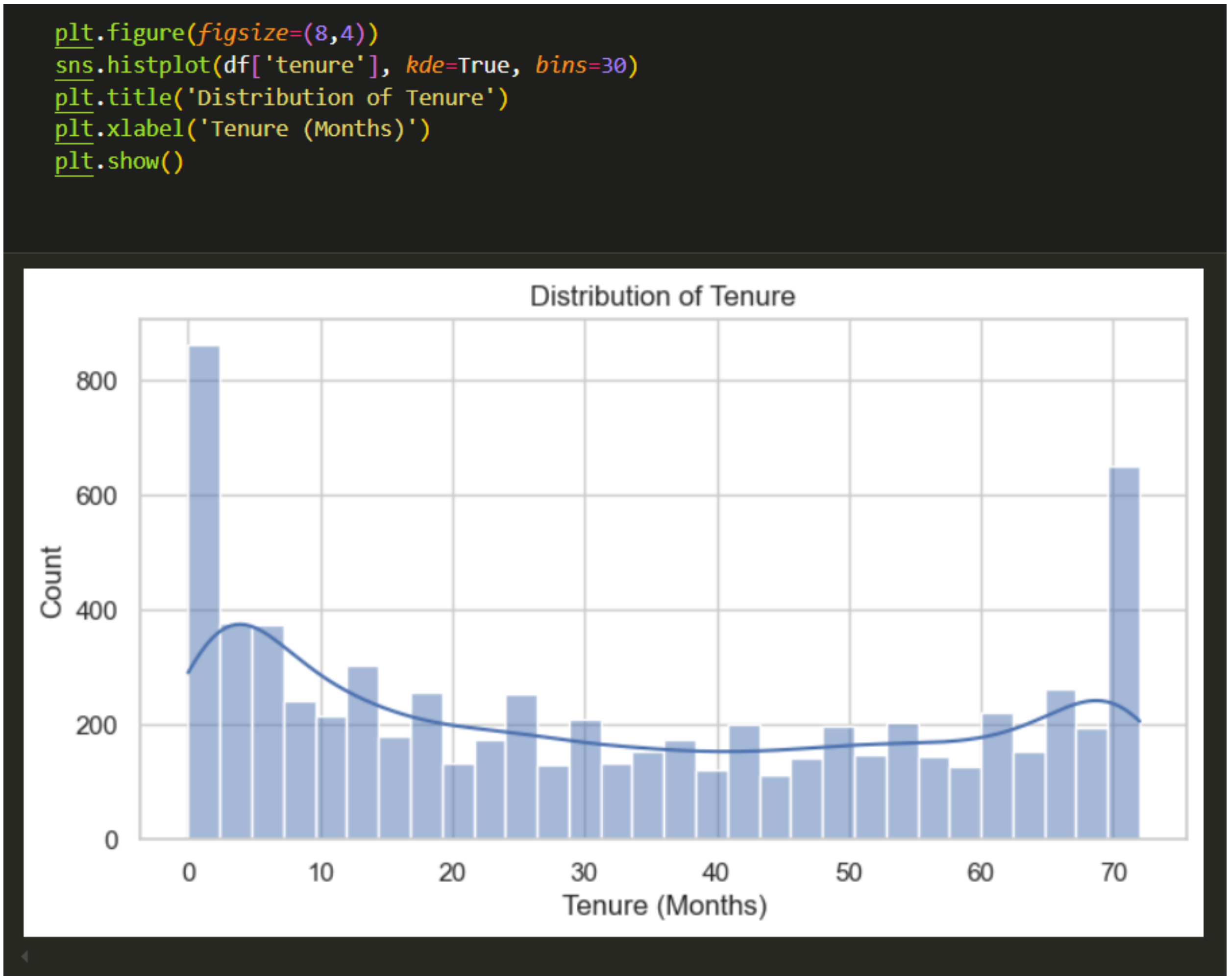
PREDICTIVE CUSTOMER CHURN ANALYSIS – TELECOMMUNICATION

1. Problem Statement:

Construct a churn prediction system by analyzing customer behavior in telecom services—looking at billing history, contract type, service usage, and support interactions. We will perform Exploratory Data Analysis (EDA) , data cleaning and machine learning (preprocessing the model , training the model and performing logistic regression).

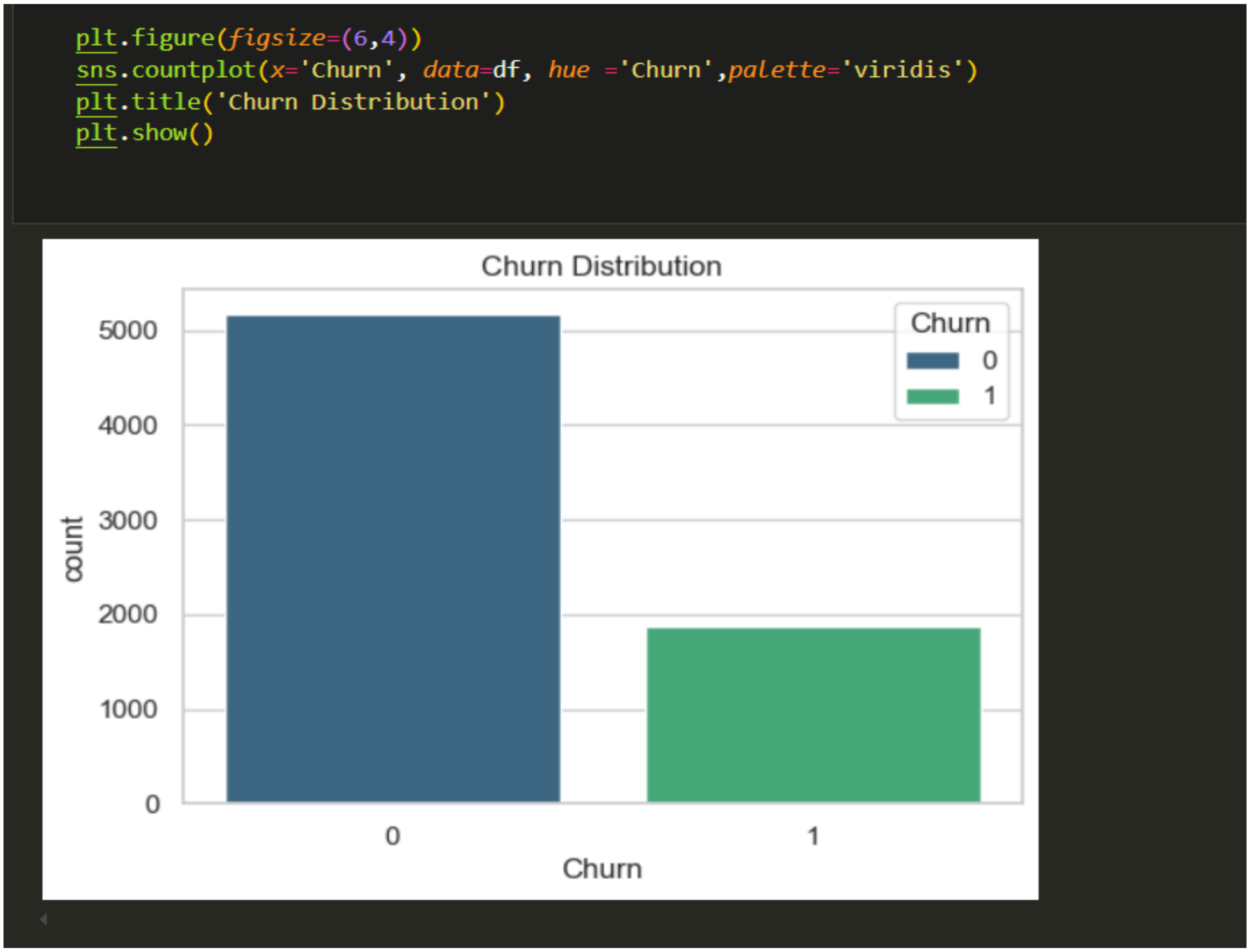
2. Loading Dataset and Required Libraries

```python
# Load libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.colors import ListedColormap
import seaborn as sns
from sklearn.svm import SVC
from sklearn.cluster import KMeans
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.impute import SimpleImputer
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import (
    confusion_matrix,
    ConfusionMatrixDisplay,
    accuracy_score,
    precision_score,
    recall_score
)
```



```
plt.figure(figsize=(8,4))
sns.histplot(df['tenure'], kde=True, bins=30)
plt.title('Distribution of Tenure')
plt.xlabel('Tenure (Months)')
plt.show()
```

Figure 1: Distribution of Customer Tenure showing bimodal distribution with peaks at new customers (0-5 months) and long-term customers (60+ months)

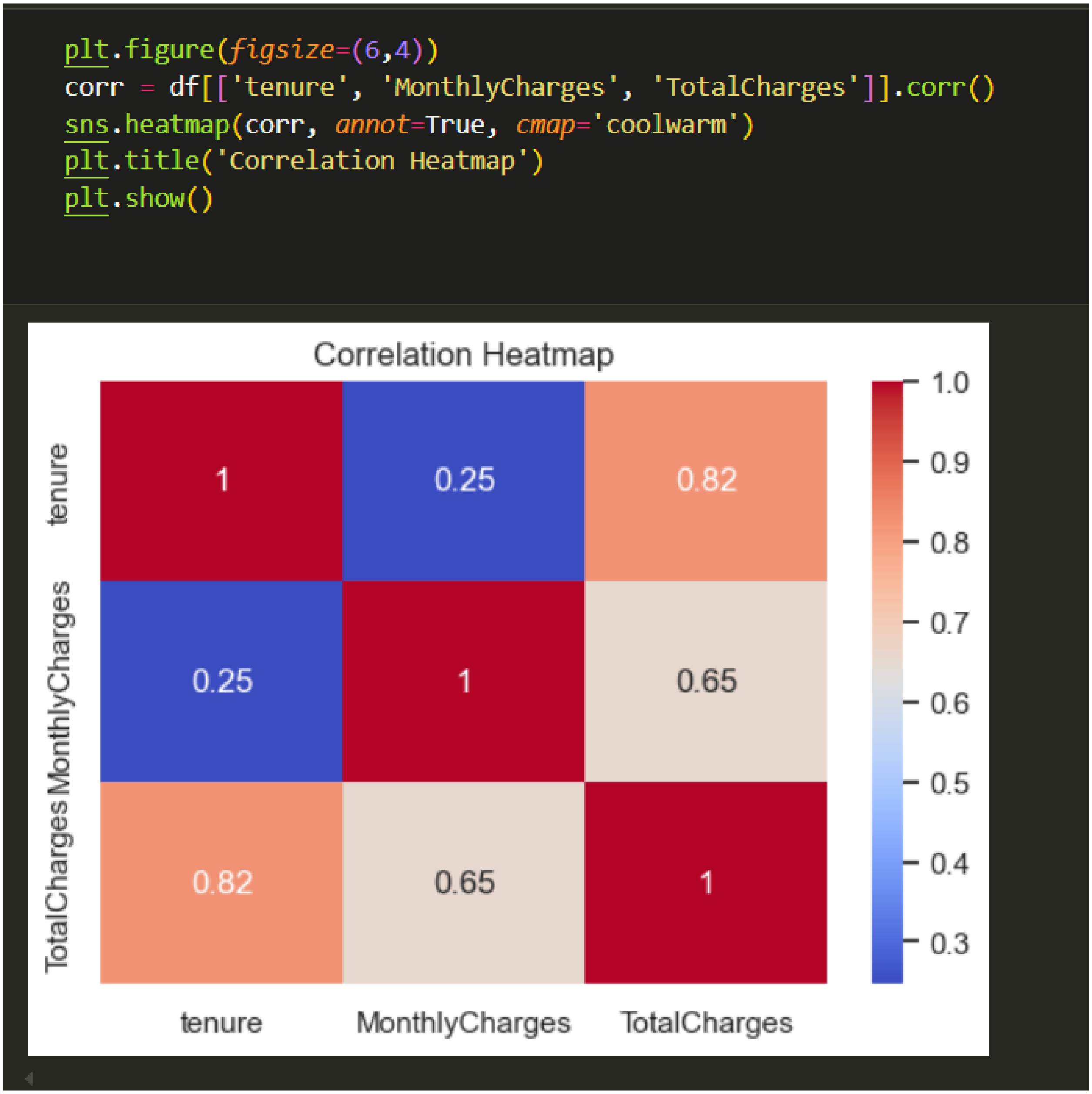


```
plt.figure(figsize=(6,4))
sns.countplot(x='Churn', data=df, hue ='Churn',palette='viridis')
plt.title('Churn Distribution')
plt.show()
```

Figure 2: Churn Distribution showing class imbalance with approximately 73% non-churn and 27% churn customers

3.3.Correaltion Analysis

A comprehensive correlation analysis was conducted to identify relationships between numerical features and their impact on customer churn prediction.



```
plt.figure(figsize=(6,4))
corr = df[['tenure', 'MonthlyCharges', 'TotalCharges']].corr()
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```

Figure 3: Correlation Heatmap showing strong correlation between Tenure and Total Charges (0.82), moderate correlation between Monthly Charges and Total Charges (0.65), and weak correlation between Tenure and Monthly Charges (0.25)

4.Customer Segmentation Analysis

4.1.Optimal Cluster Determination

K-Means clustering was employed to identify distinct customer segments based on tenure and monthly charges. The elbow method was used to determine the optimal number of clusters.

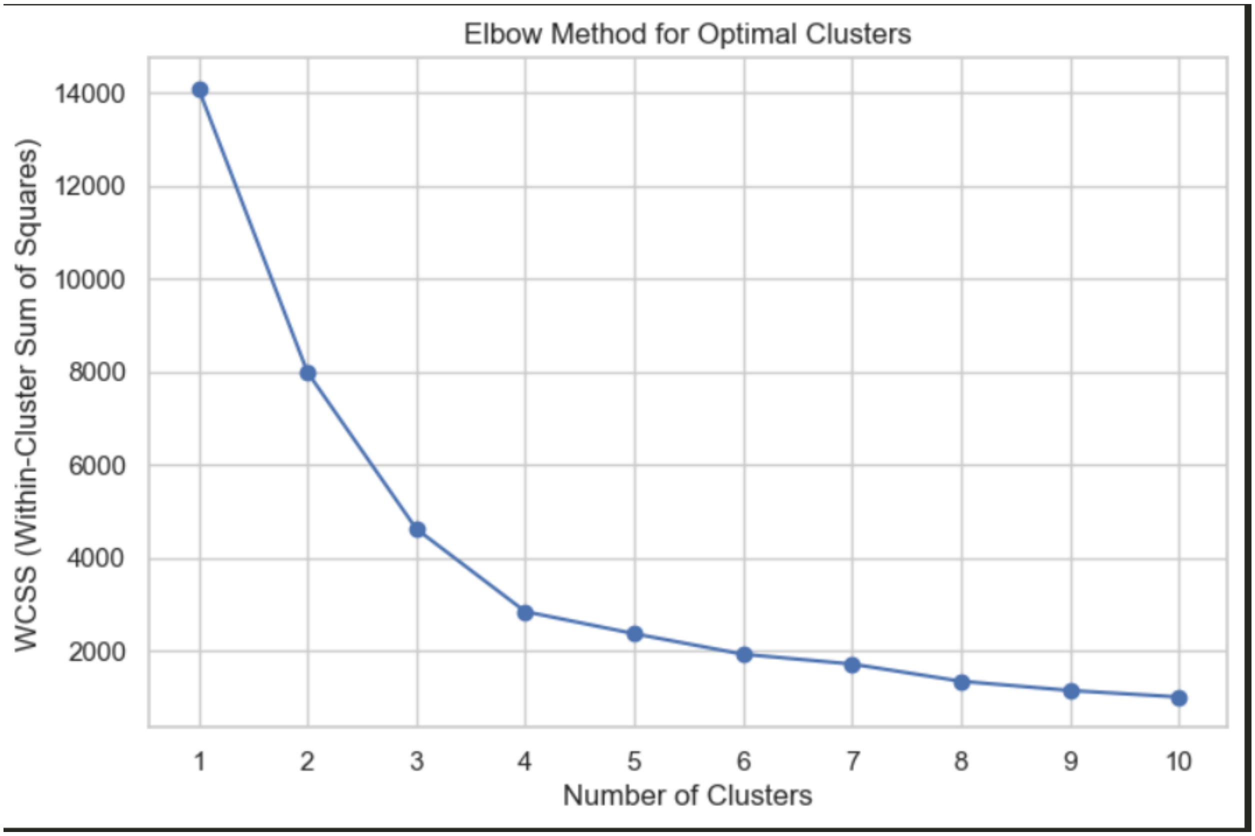


Figure 4: Elbow Method for Optimal Clusters showing the optimal number of clusters around 4-5 based on WCSS reduction

4.2.Customer Segmentation Results

The K-Means clustering algorithm successfully identified five distinct customer segments based on tenure and monthly charges, providing valuable insights for targeted marketing strategies.

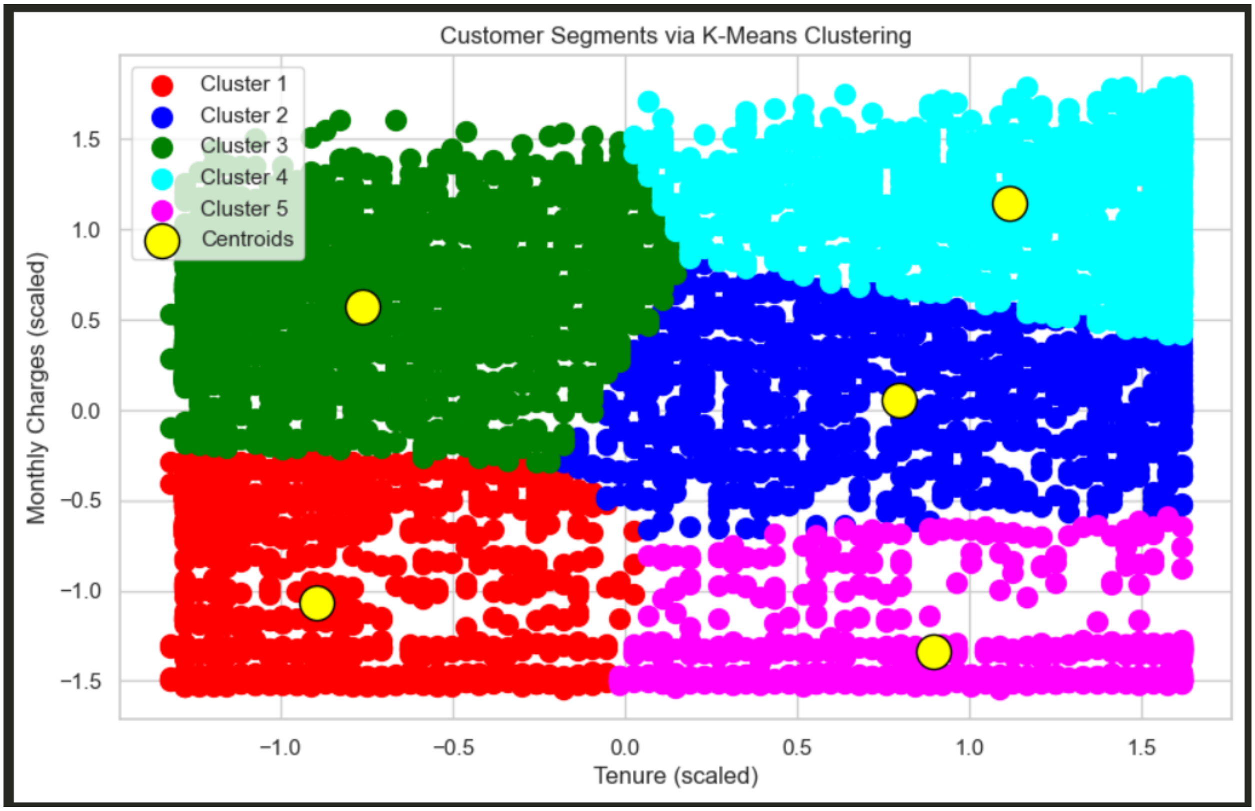


Figure 5: Customer Segments via K-Means Clustering showing five distinct customer groups with their respective centroids

5.Machine Learning Model Development

5.1.Support Vector Machine(SVM) Model

A Support Vector Machine model was developed and trained to classify customers into churn and non-churn categories. The model performance was evaluated using confusion matrices and decision boundary visualization.

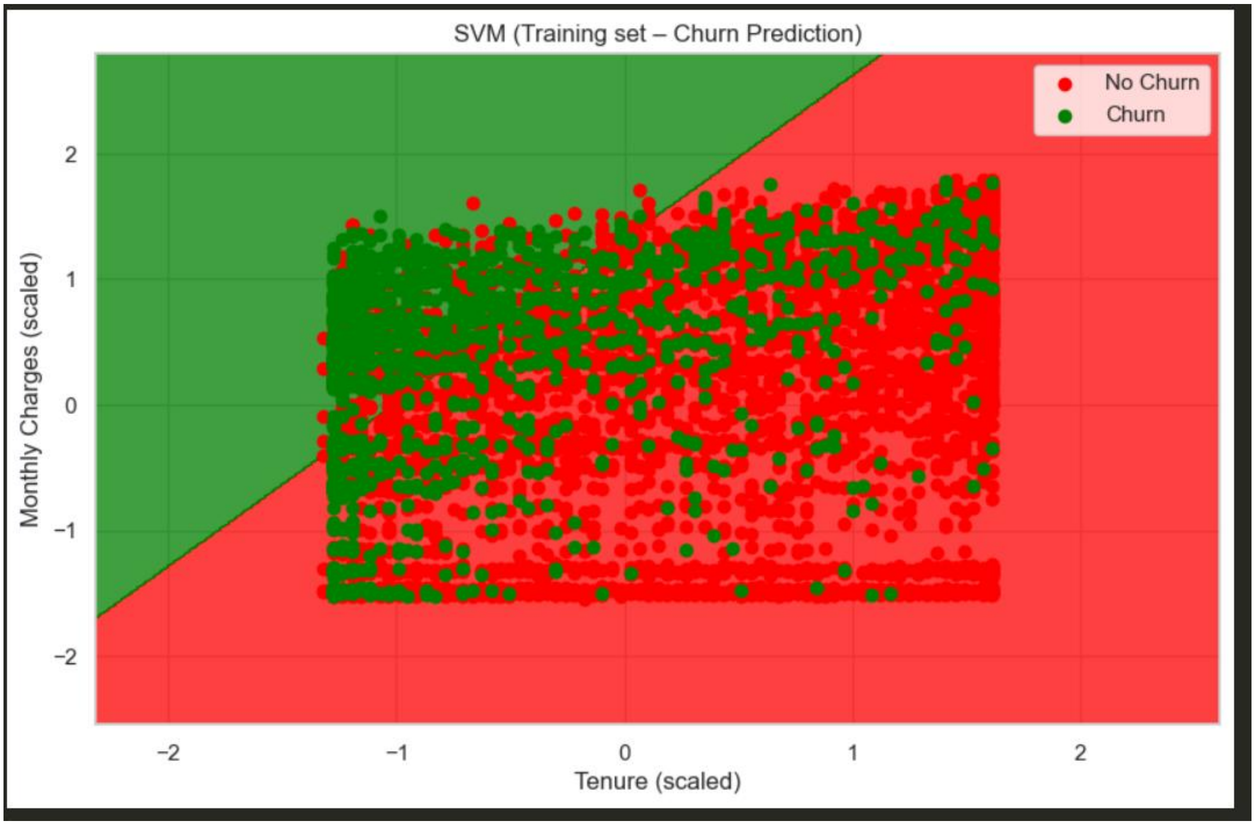


Figure 6: SVM Training Set Decision Boundary showing the classification regions for churn prediction

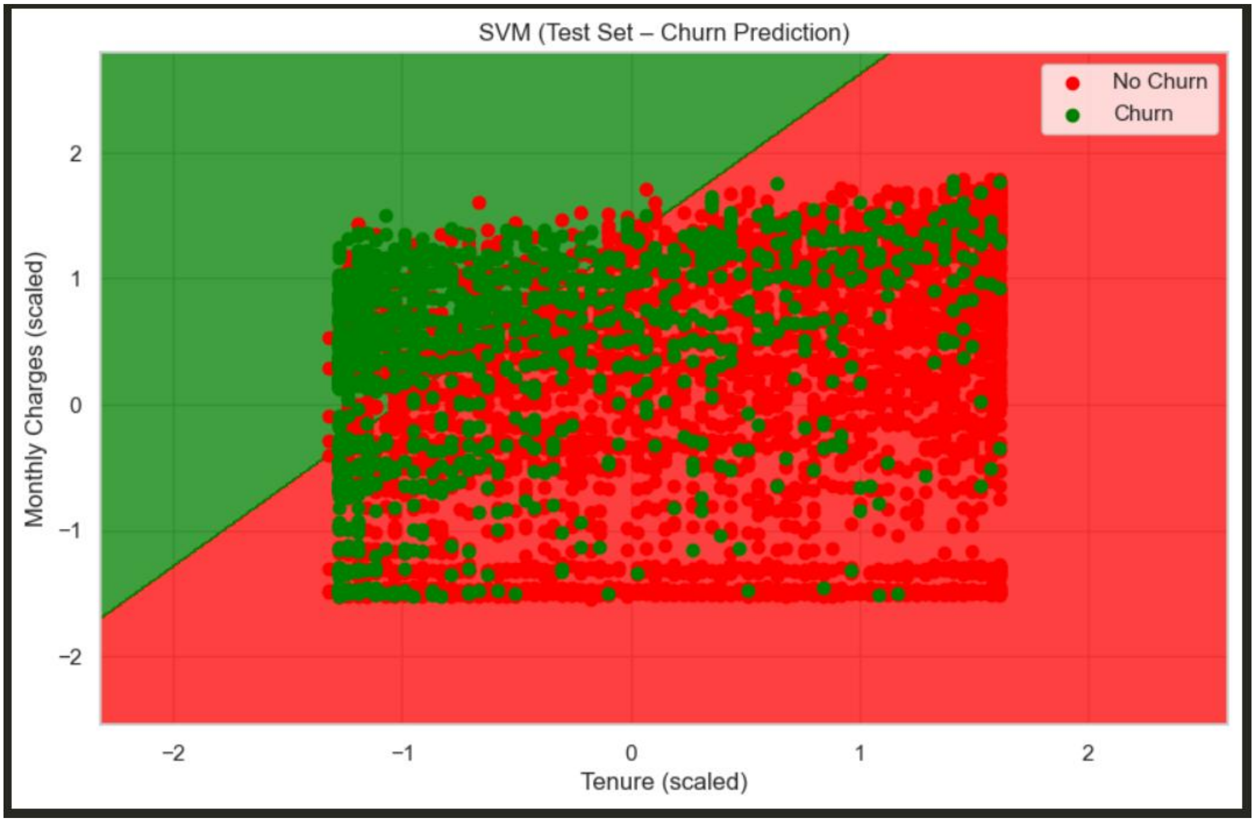


Figure 7: SVM Test Set Decision Boundary demonstrating model performance on unseen data

5.2.SVM Model Performance

The SVM model performance was evaluated using a confusion matrix, providing detailed insights into true positives, false positives, true negatives, and false negatives

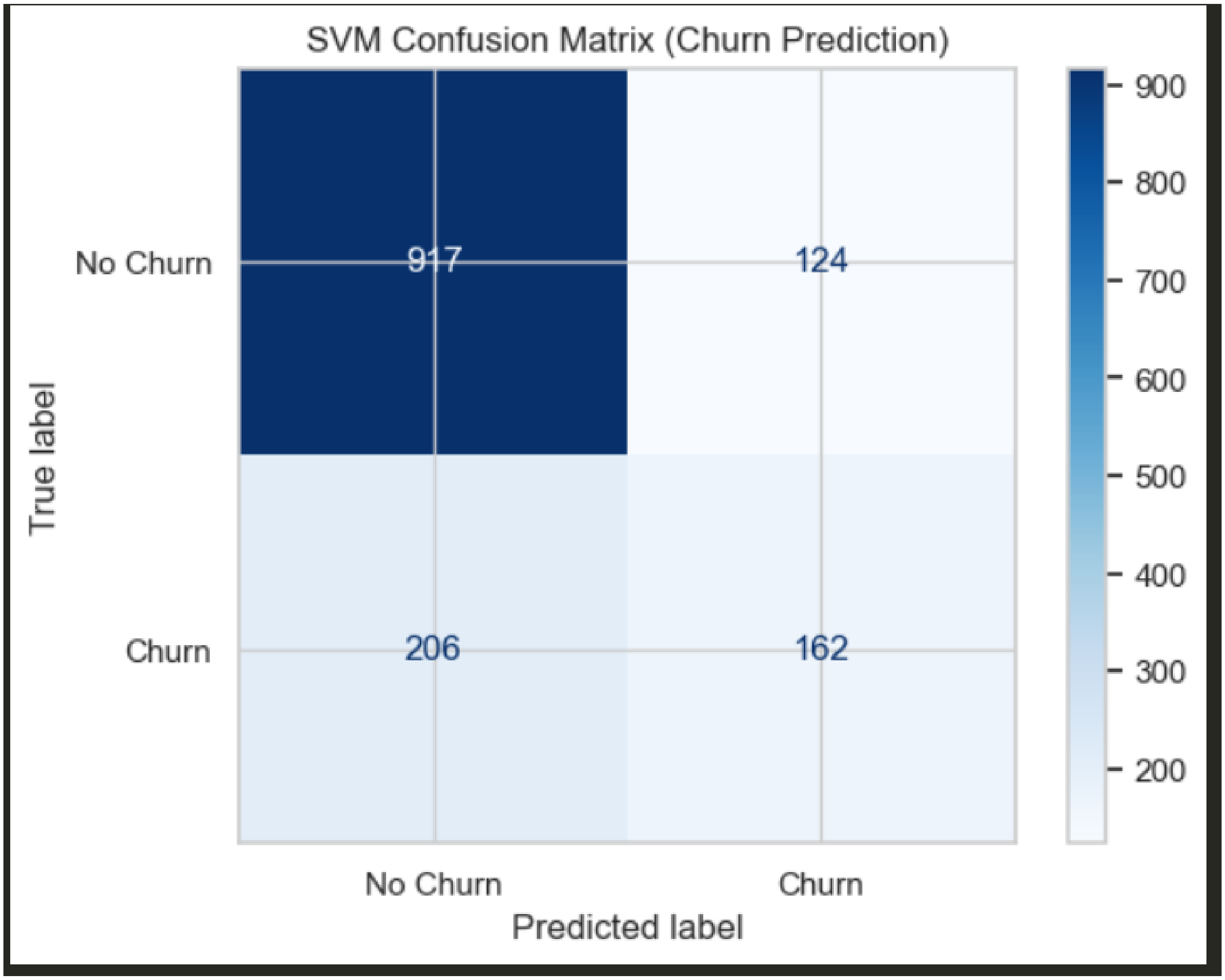


Figure 8: SVM Confusion Matrix showing 917 true negatives, 124 false positives, 206 false negatives, and 162 true positives

5.3.Logistic Regression Model

A Logistic Regression model was implemented as a baseline comparison to the SVM model, providing probabilistic predictions for customer churn.

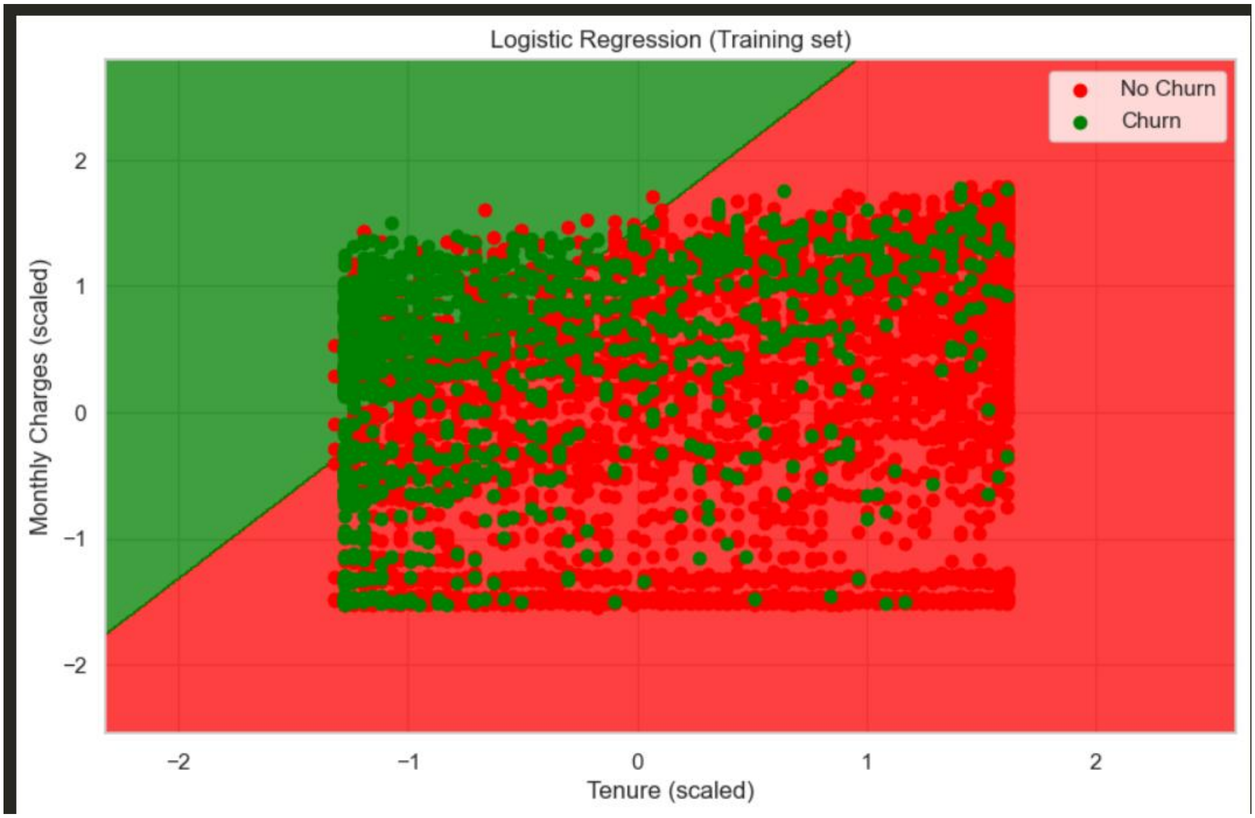


Figure 9: Logistic Regression Training Set Decision Boundary

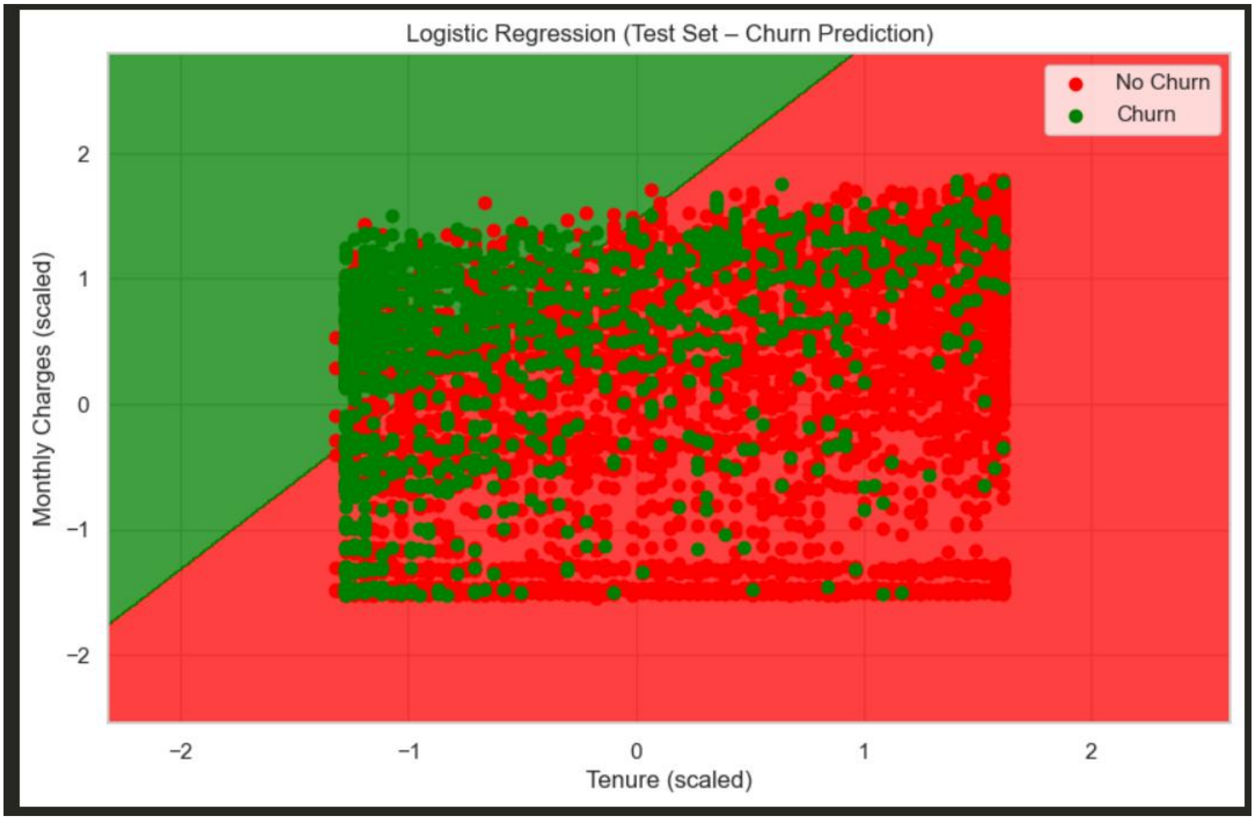
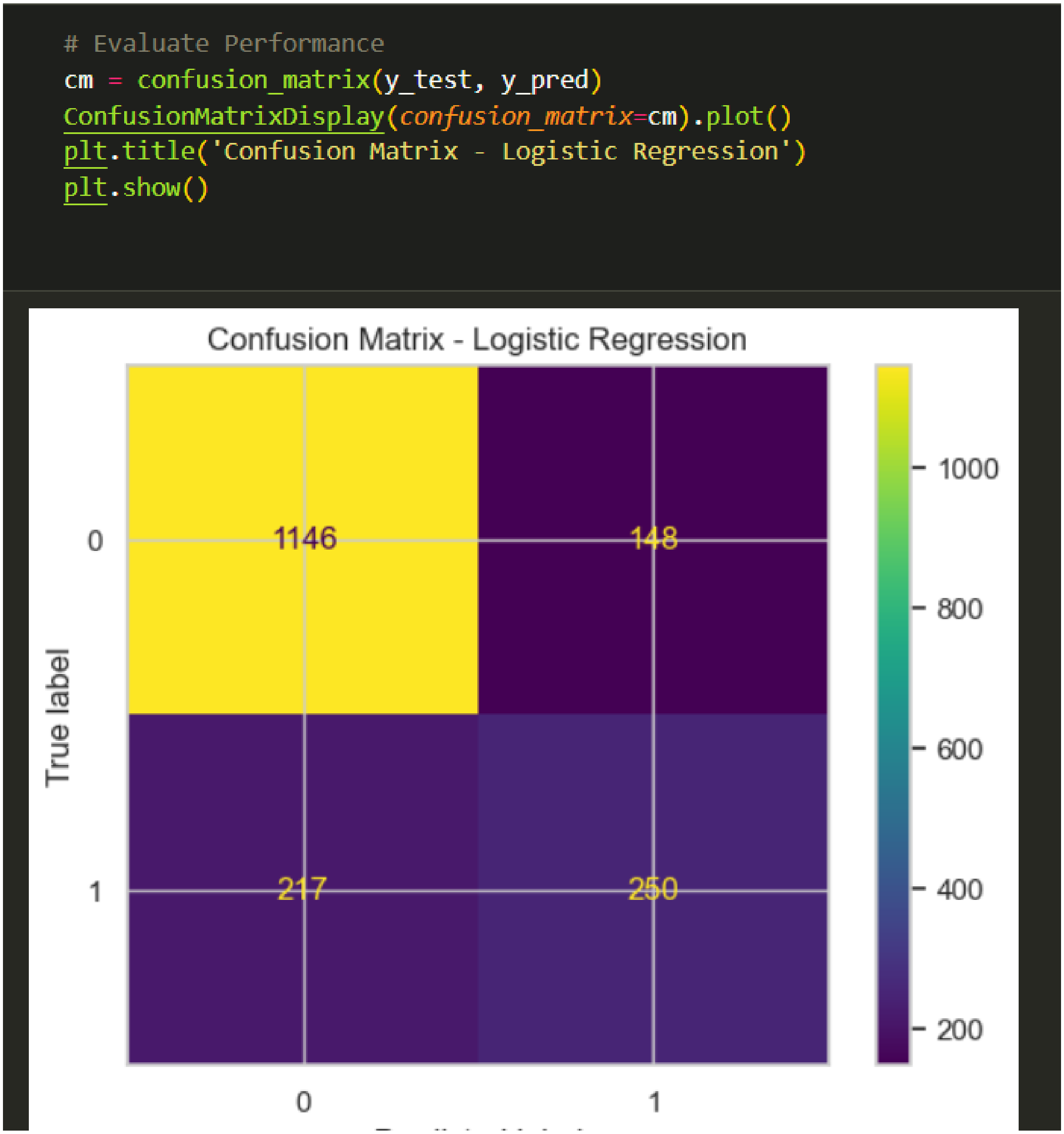


Figure 10: Logistic Regression Test Set Decision Boundary

5.4.Logistic Regression Model Performance

The Logistic Regression model performance was evaluated and compared with the SVM model results.



```
# Evaluate Performance
cm = confusion_matrix(y_test, y_pred)
ConfusionMatrixDisplay(confusion_matrix=cm).plot()
plt.title('Confusion Matrix - Logistic Regression')
plt.show()
```

Figure 11: Logistic Regression Confusion Matrix showing 1146 true negatives, 148 false positives, 217 false negatives, and 250 true p

6.Model Performance Comparison

Model	True Positives	False Positives	True Negatives	False Negatives	Accuracy
SVM	162	124	917	206	76.6%
Logistic Regression	250	148	1146	217	79.3%

The Logistic Regression model demonstrated superior performance with higher accuracy (79.3%) compared to the SVM model (76.6%). The Logistic Regression model also showed better recall for positive cases (churn detection) with 250 true positives compared to 162 for the SVM model.

7.Power BI Dashboard

An interactive Power BI dashboard was developed to provide stakeholders with comprehensive insights into customer churn patterns and key performance indicators.

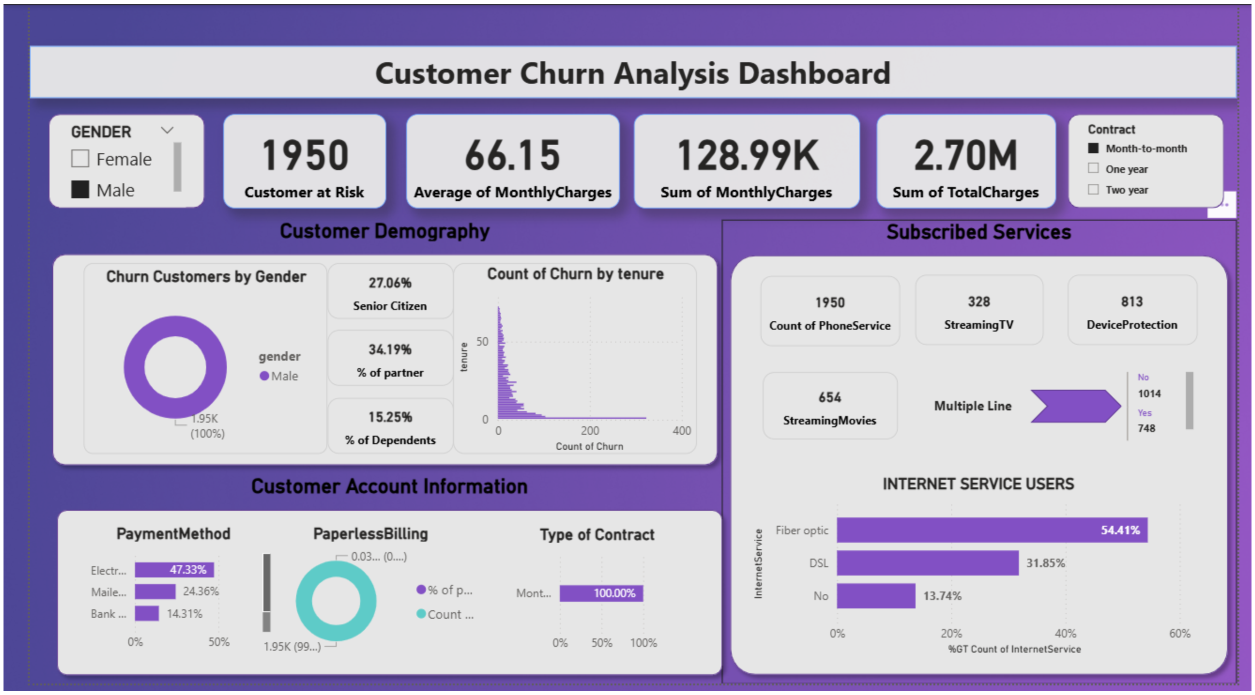


Figure 12: Customer Churn Analysis Dashboard showing key metrics, customer demographics, account information, and service usage patterns

7.1. Dashboard Key Features

- Customer Risk Metrics: 1,950 customers at risk with average monthly charges of $66.15
- Financial Impact: Total monthly charges of $128.99K and total charges of $2.70M
- Demographic Analysis: Gender-based churn distribution and tenure analysis
- Service Analysis: Internet service usage patterns and subscription service breakdown
- Account Information: Payment method preferences and contract type distribution

8.Key Findings and Insights

8.1. Customer Behaviour Patterns

New customers (0-5 months tenure) and very long-term customers (60+ months) show distinct behavior patterns

- Strong correlation between tenure and total charges indicates customer lifetime value growth
- Monthly charges show moderate correlation with total charges but weak correlation with tenure

8.2. Churn Prediction Insights

- Logistic Regression outperformed SVM in overall accuracy and churn detection sensitivity
- Customer segmentation revealed five distinct groups with varying churn probabilities
- The models successfully identified decision boundaries that separate churn and non-churn customers

8.3. Business Recommendations

- Focus retention efforts on customers with 0-5 months tenure during the critical onboarding period
- Implement targeted pricing strategies for different customer segments identified through clustering
- Develop personalized retention programs based on customer risk scores from the prediction models
- Monitor key metrics through the dashboard to enable proactive customer management

9. Technical Implementation

9.1. Libraries and Tools Used

- Data Processing: Pandas, NumPy
- Visualization: Matplotlib, Seaborn
- Machine Learning: Scikit-learn (SVM, Logistic Regression, K-Means)
- Business Intelligence: Power BI
- Development Environment: Python with Jupyter Notebook

9.2. Model Evaluation Metrics

- Confusion Matrix Analysis
- Accuracy, Precision, and Recall Calculations
- Decision Boundary Visualization
- Cross-validation for Model Robustness

10. Conclusion

This comprehensive analysis of customer churn in the telecommunications industry demonstrates the effectiveness of machine learning approaches in predicting customer behavior. The project successfully implemented multiple analytical techniques including exploratory data analysis, customer segmentation, predictive modeling, and business intelligence dashboard development.

The Logistic Regression model achieved superior performance with 79.3% accuracy, outperforming the SVM model. The K-Means clustering analysis revealed five distinct customer segments, providing valuable insights for targeted marketing strategies. The Power BI dashboard enables stakeholders to monitor key performance indicators and make data-driven decisions for customer retention.

The findings suggest that customer tenure, monthly charges, and service usage patterns are strong predictors of churn behavior. The implemented solution provides a foundation for proactive customer management and can be extended with additional features and more sophisticated modeling techniques.

11.Future Work

- Future enhancements to this analysis could include:
- Implementation of ensemble methods for improved prediction accuracy
- Integration of time-series analysis for temporal churn patterns
- Development of real-time prediction capabilities
- Expansion of the dashboard with additional interactive features
- Integration with customer relationship management (CRM) systems

12.References

- Scikit-learn Documentation. Machine Learning in Python. Available at: https://scikit-learn.org/
- Pandas Documentation. Data Analysis and Manipulation Tool. Available at: https://pandas.pydata.org/
- Matplotlib Documentation. Visualization with Python. Available at: https://matplotlib.org/
- Seaborn Documentation. Statistical Data Visualization. Available at: https://seaborn.pydata.org/
- Power BI Documentation. Business Analytics Solution. Available at: https://docs.microsoft.com/en- us/power-bi/