Capstone Final Report

Telecom Churn Rate

Introduction

Why do customers leave, and can your business afford to let your customers slip away?

The telecom industry is highly competitive, with companies constantly vying for customer loyalty. A key metric that helps evaluate the health of a telecom business is the churn rate—the percentage of customers who leave a service over a given period. Understanding churn is critical for telecom companies, as high churn rates often indicate dissatisfaction, unmet customer needs, or better offers from competitors. Using this dataset, we were able to identify the factors driving churn and devise strategies to improve customer retention, reduce acquisition costs, and boost overall profitability.

Problem Statement

How should a Switzerland mid-scale wireless service company lower churn rate by the end of the year in order to (a) achieve at least 70% predictive accuracy in determining which customers will churn and (b) achieve a 10% decrease in churn rate?

Data Wrangling

The raw dataset from Kaggle contained 7043 rows and 23 columns. It offered a sufficiently large sample size, which is ideal for developing predictive models. A larger dataset typically helps create more accurate models and minimizes overfitting, leading to more reliable insights. It is large

enough to identify trends, patterns, and relationships in the data, which are essential for understanding customer behavior and churn.

When working with datasets, missing values are a common challenge that can affect the quality and performance of your models. To address this, one approach I took is to replace the missing values with the statistical mean of the respective column. Replacing missing values with the mean ensures that the data remains consistent without introducing bias or significant distortion.

Exploratory Data Analysis

Analyzing the churn rate dataset often involves exploring various factors that might influence customer churn, and one critical question to consider is whether pricing plays a role in churn. By exploring relationships between customer characteristics, spending, and churn, companies can gain valuable insights into why customers leave. In particular, variables such as contract type, tenure, senior citizen, monthly and total charges were analyzed to identify correlations with churn.

The first factor I was interested in was “tenure”. In figure 1 I wanted to observe the length of time customers has been with this company or subscribed to a service

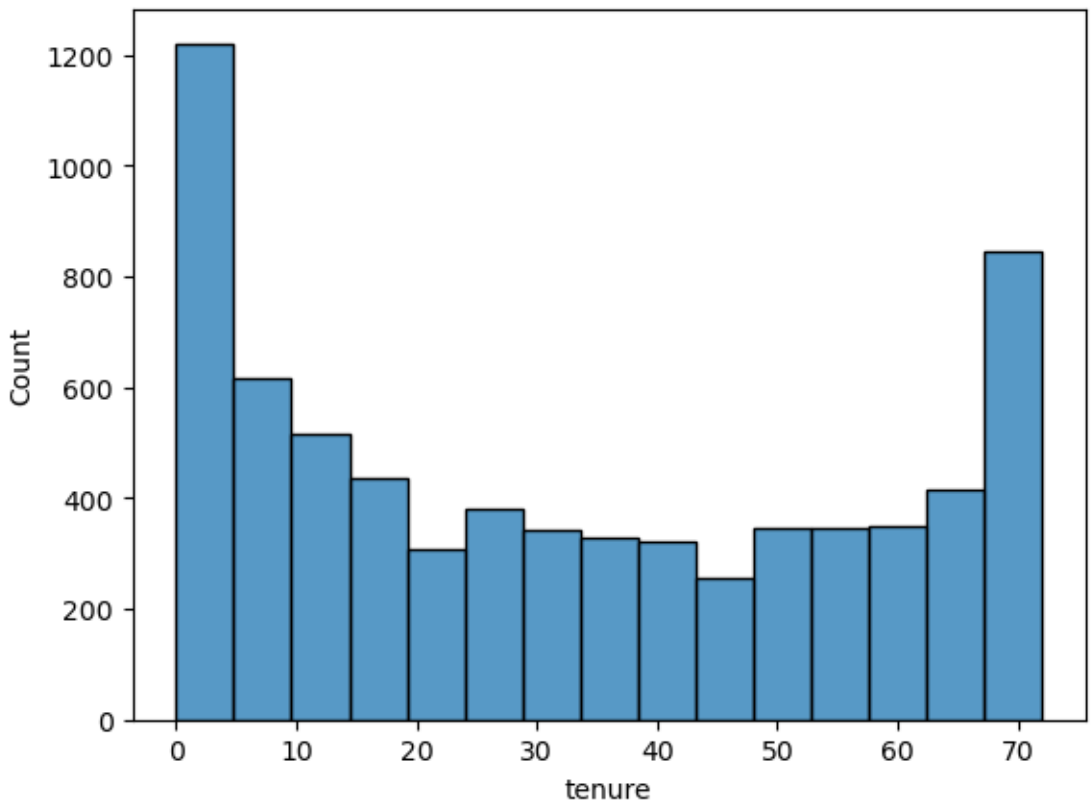

Figure 1. Histogram of Tenure (Months) vs count

The histogram in Figure 1 indicated that most customers either leave in the 1st month or stay for more than 6 years.

Other factors that were observed were senior citizen, term of contract, churn, monthly charges, and total charges.

Senior citizen were shown to represent most of the customers taken up 83%.

	Senior Citizen
0	5885
1	1141

Name: count, dtype: int64

There were also shown to provide only 3 types of contracts:

Contract

Month-to-month 3858

Two year 1695

One year 1473

Name: count, dtype: int64

From this data most customers decided to go with Month-to-month contracts.

Out of 7,026 customers, 5,166 remained with the company, while 1,860 churned, resulting in a churn rate of 26%. This means that 26% of customers canceled their subscriptions, which is not significantly higher than the typical churn rate observed in the telecom industry. According to studies, the average monthly churn rate for telecom companies is around 1.9% per month across major carriers. When this is converted to an annual churn rate, it would typically range between 20% to 40% annually, depending on the specific market conditions and company strategies.

Now to determine if pricing plays a role in churn, I viewed the relationship that Monthly Charges (Figure 2) and Total Charges (Figure 3) have with Churn rate.

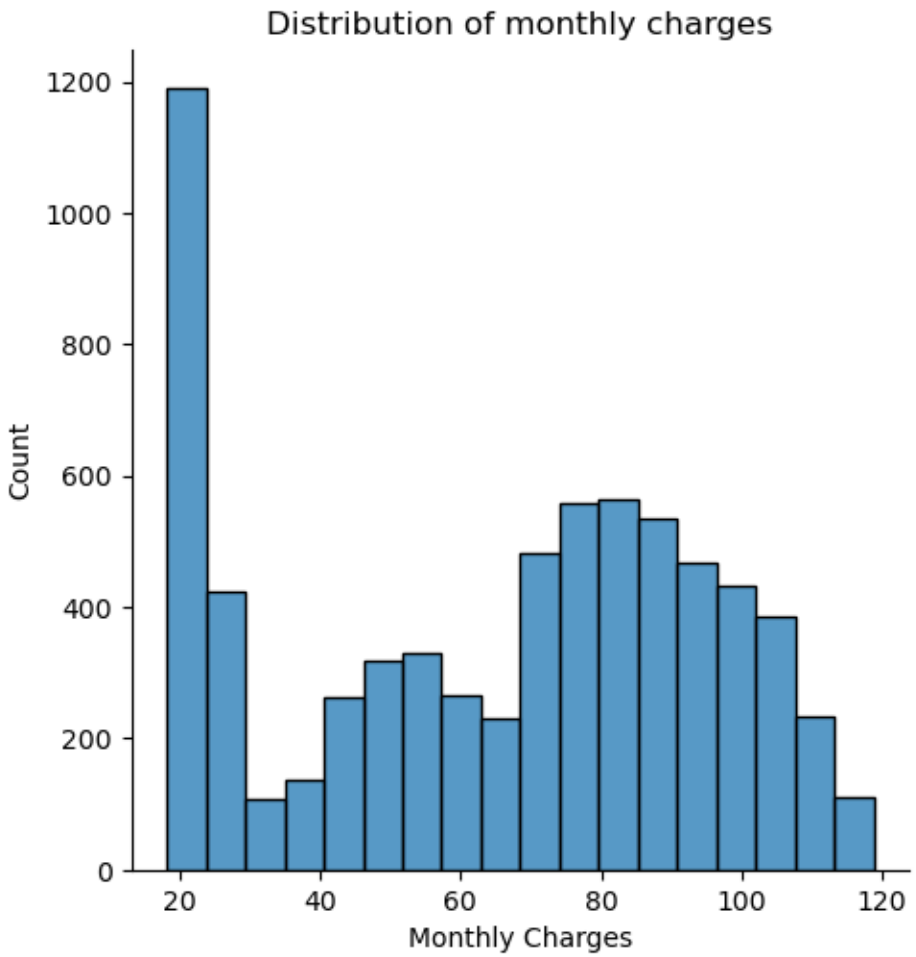

Figure 2. Histogram of Monthly Charges ($) vs Count

Figure 2 indicates that most consumers tend to purchase products priced around $19 or $20, making this price range a key area with significant marketing potential.

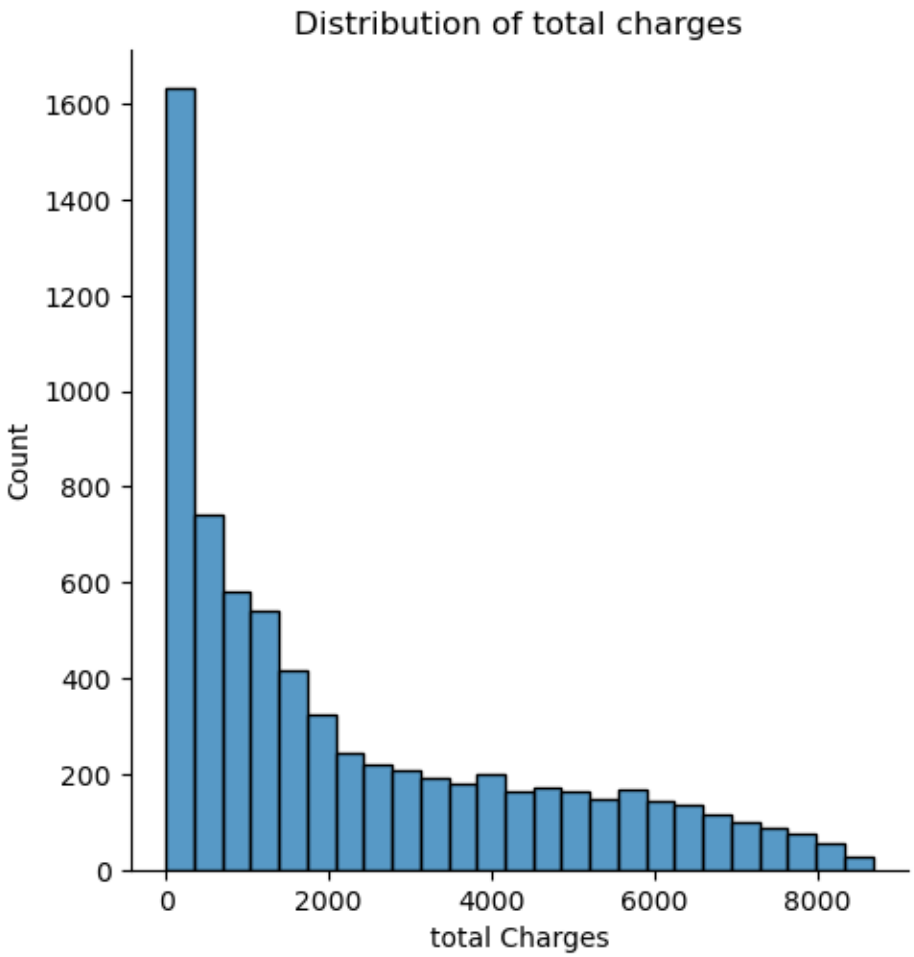

Figure 3. Histogram of Total Charges ($) vs Count

The histogram in Figure 3 indicates as total charges increase, the count of consumers or transactions tends to decrease. This suggests that fewer customers are willing or able to spend at higher price points, indicating a potential shift in customer behavior or a market segment that is more sensitive to price increases.

Model Selection

Determining which model to use can be challenging, and several struggles often arise in this process. With the data given, I know KNN performs particularly well on datasets with moderate sizes, such as a churn dataset with several thousand rows. KNN is relatively simple to understand and

implement. It works by classifying a new data point based on the majority class (or average in the case of regression) of its k nearest neighbors in the feature space. For predicting the relationship between Monthly Charges, Tenure, and Churn, I chose K-Nearest Neighbors (KNN) as one of the model options. After splitting the data into training and testing sets, I initialized the KNN classifier and trained it on the training data. The model achieved an accuracy of 76% on the test set. To visualize this, I used a ROC curve In Figure 4. and a Confusion Matrix in Figure 5. to capture the True Positives Rate (TPR) and the False Positive Rate (FPR) of churn.

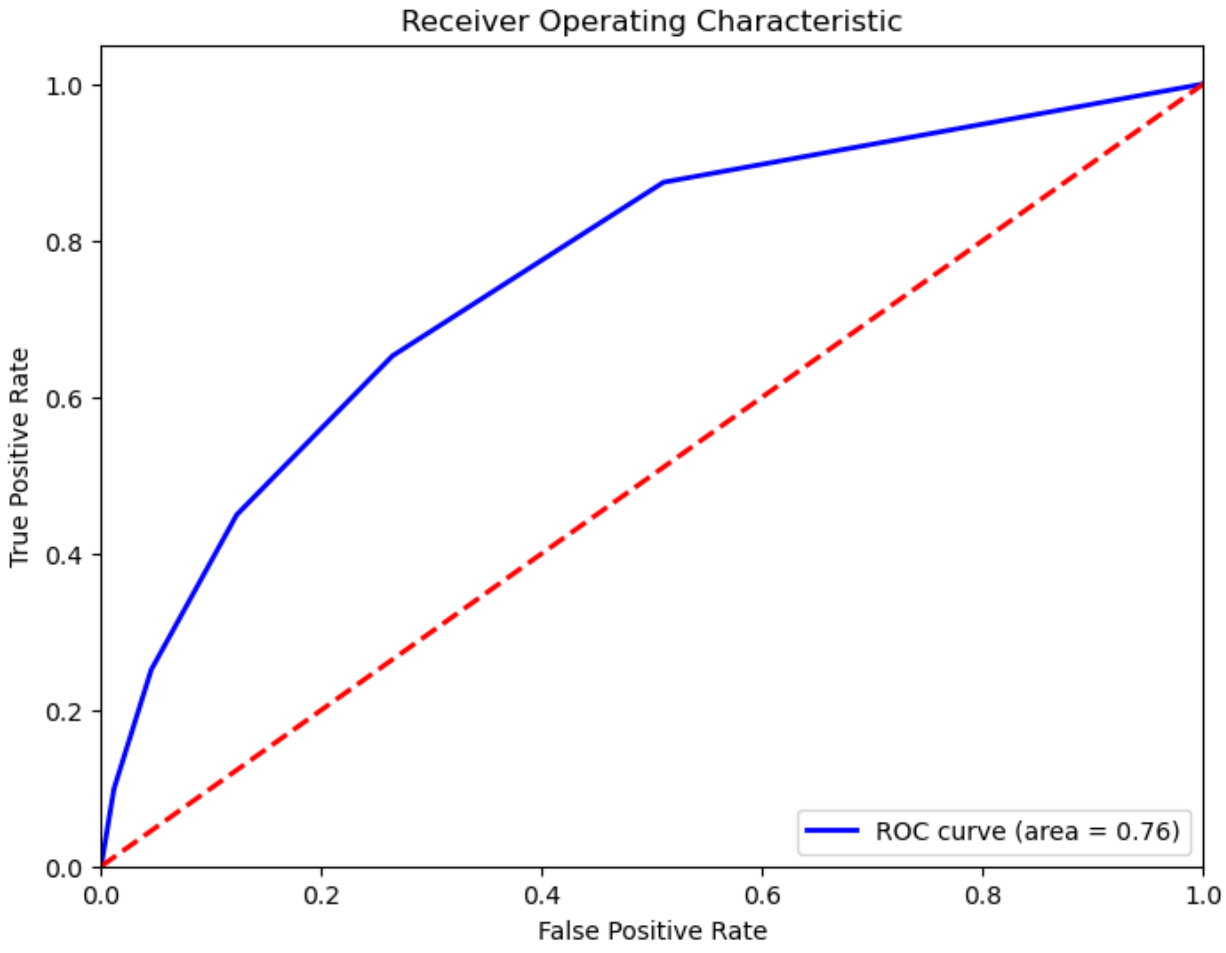

Figure 4. ROC Curve of the False Positive Rate vs True Positives Rate

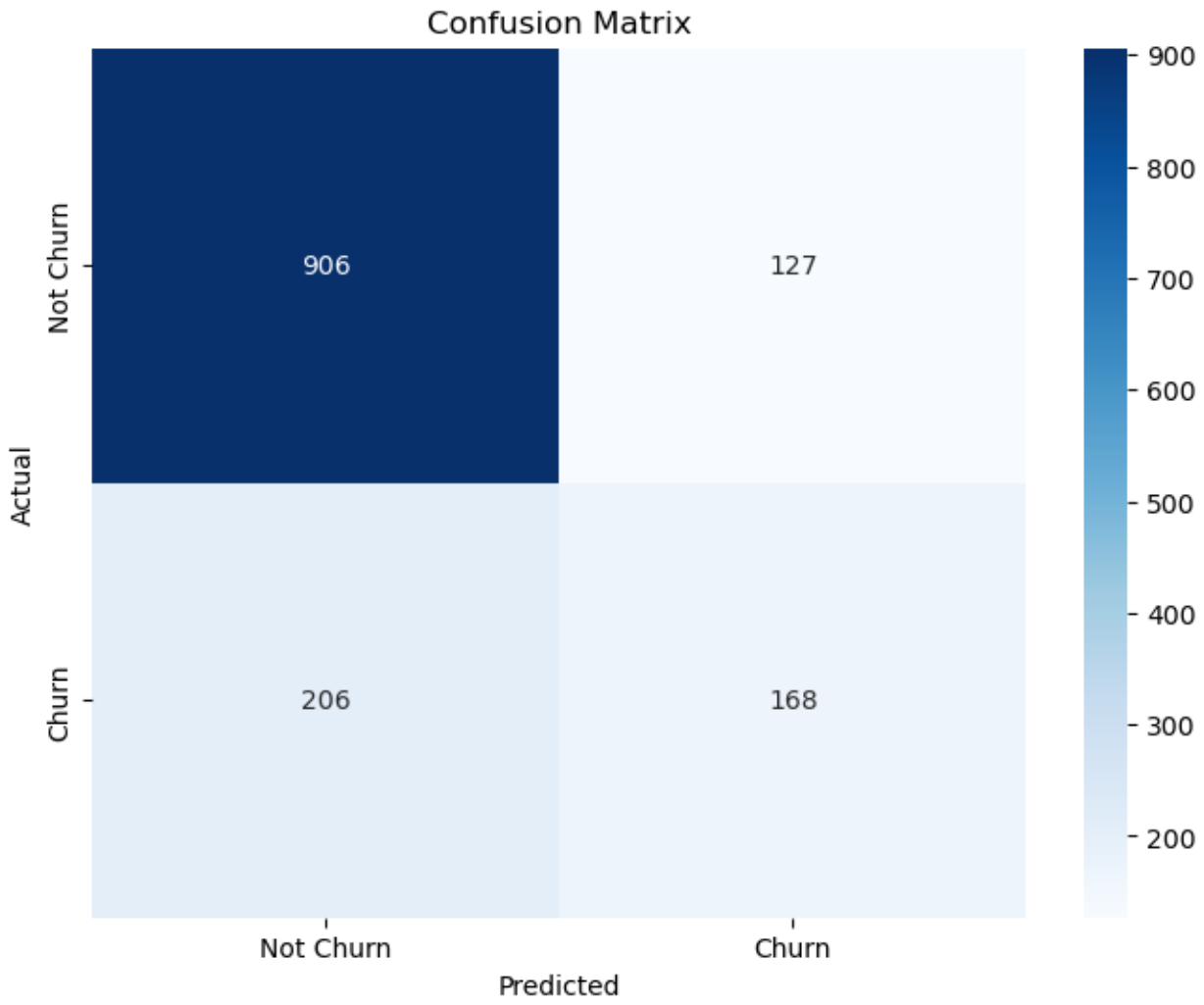

Figure 5. Confusion Matrix of Predicted vs Actual

Figure 5 heatmap displays counts of actual versus predicted classifications.

The other model that I used was linear regression, because it provides clear insights into the relationship between independent variables (customer features like monthly charges, tenure) and the dependent variable. I also wanted to see if there was a linear relationship between input variables and the target variable.

After creating the regression model and fitting the data, the model obtained higher predictive values than actual values. The large differences between the predicted and actual values suggest that the model might not be capturing the true patterns in the data. Figure 6. Shows the Tenure vs Total charges.

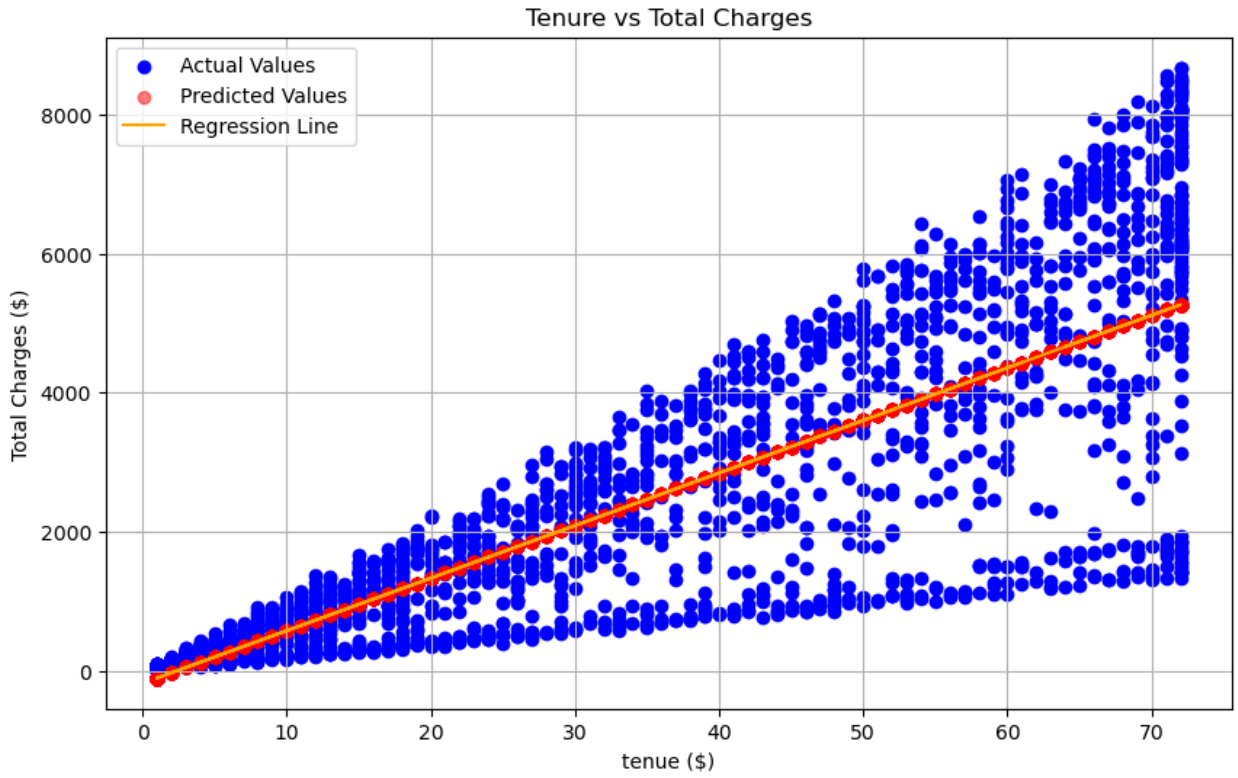

Figure.6 regression line plot of Tenure vs Total Charges ($)

Figure 6 does show a positive relationship between tenure and total charges. This may happen because a s a customer remains with the company for a longer period, their total charges tend to increase, because they are continuously paying for the services.

The R-squared and the Root Mean Squared Error were:

R^2: 0.6908213475789509

RMSE: 1266.5826559770207

An R^2 value of 0.69 is generally considered to be good, especially for complex, real-world datasets.

The RMSE of 1266.58 suggests that the model's average prediction error is around 1266.58 units. Depending on the scale of the target variable, this error may be considered relatively high and could be a potential area for improvement.

Takeaways

Out of the two models, K-Nearest Neighbors (KNN) performed better, achieving a 76% accuracy, which outperforms the linear regression model, and reached our 70% target margin. While linear regression demonstrated a solid coefficient of determination (R^2), indicating that the independent variables explain a good portion of the variability in the target variable, the RMSE (Root Mean Squared Error) of 1266.58 suggests that the model's predictions have a relatively high error margin. This higher RMSE indicates there is room for improvement in the linear regression model's predictive performance, particularly in reducing the prediction errors.

When examining the relationship between monthly charges and Total Charges, churned customers appear to have a higher average spend than non-churned customers, suggesting that those who spend less are more likely to remain with the service. So, for this Switzerland Telecom company my recommendation to achieve a 10% decrease in churn rate, is to apply new contracts prices that is responsible for at least 80% of our work requests. Marketing practices should be implemented for the senior citizens no later than January 2023. Senior citizens are the largest group representing 83% of the customers. My last recommendation would me to offer targeted loyalty programs and discounts to long-term customers, especially those who have been with the company for over a year. By rewarding customer loyalty, particularly among high-value accounts, the company can reduce churn and

increase customer satisfaction at the 6 year mark which is shown to have a high churn.

Future Research

Future research based on this telecom dataset could focus on refining the performance of both K-Nearest Neighbors (KNN) and Linear Regression to further improve churn prediction accuracy and overall model effectiveness. One avenue for improvement would be exploring hyperparameter tuning for KNN, such as optimizing the number of neighbors (k) or experimenting with different distance metrics to better capture the underlying patterns in the data.

For Linear Regression, a deeper exploration of non-linear relationships could be beneficial, potentially incorporating polynomial features or transforming variables to better model complex interactions between predictors. Additionally, addressing the model's high RMSE could involve experimenting with more sophisticated regression techniques, like regularized regression (Ridge or Lasso), or applying more advanced machine learning algorithms such as random forests or gradient boosting, which might provide more accurate predictions by capturing non-linearities and interactions.