Name: Muhammad Ibrahim

Customer Churn Prediction Report

**Introduction**

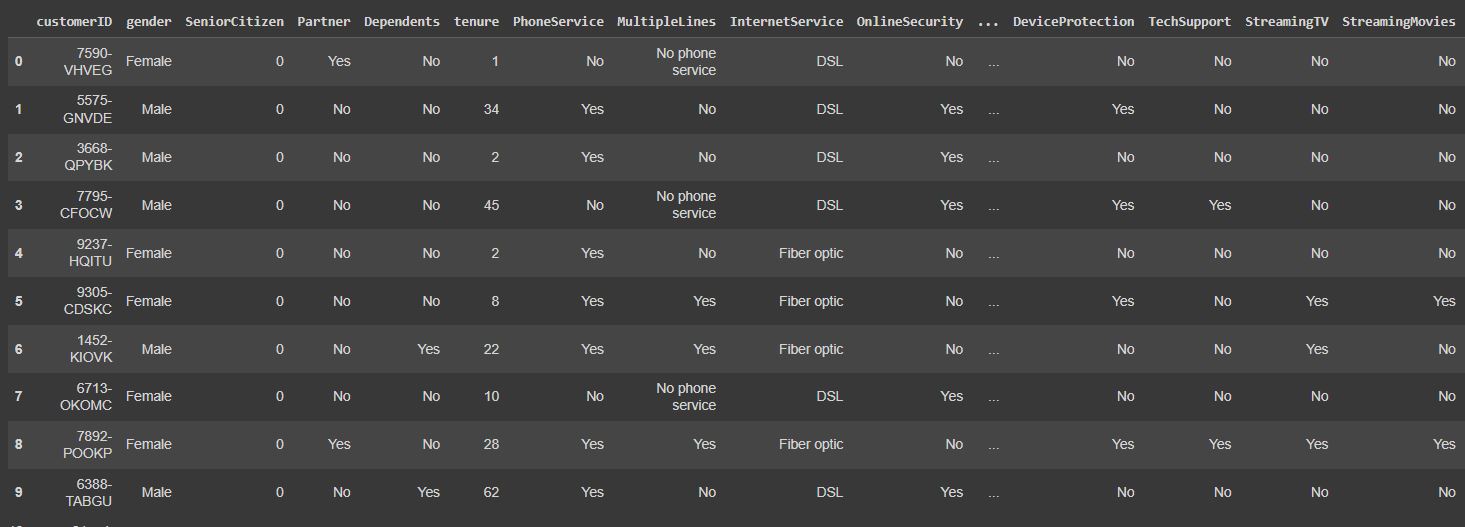
This report presents the analysis and prediction of customer churn for a telecommunications company. The goal is to develop a model that can accurately predict whether a customer is likely to churn or not. The dataset used for this analysis is the Telco Customer Churn dataset.

**Dataset Overview**

The Telco Customer Churn dataset contains information about various factors that may contribute to customer churn, such as customer demographics, services subscribed, and billing information. The dataset includes the following columns:

* gender: The gender of the customer (categorical: "Male" or "Female").
* SeniorCitizen: Indicates whether the customer is a senior citizen or not (categorical: 1 or 0).
* Partner: Indicates whether the customer has a partner or not (categorical: "Yes" or "No").
* Dependents: Indicates whether the customer has dependents or not (categorical: "Yes" or "No").
* tenure: The number of months the customer has stayed with the company (numerical).
* PhoneService: Indicates whether the customer has a phone service or not (categorical: "Yes" or "No").
* MultipleLines: Indicates whether the customer has multiple lines or not (categorical: "Yes", "No", or "No phone service").
* InternetService: The type of internet service the customer has subscribed to (categorical: "DSL", "Fiber optic", or "No").
* OnlineSecurity: Indicates whether the customer has online security or not (categorical: "Yes", "No", or "No internet service").
* OnlineBackup: Indicates whether the customer has online backup or not (categorical: "Yes", "No", or "No internet service").
* DeviceProtection: Indicates whether the customer has device protection or not (categorical: "Yes", "No", or "No internet service").
* TechSupport: Indicates whether the customer has tech support or not (categorical: "Yes", "No", or "No internet service").
* StreamingTV: Indicates whether the customer has streaming TV or not (categorical: "Yes", "No", or "No internet service").
* StreamingMovies: Indicates whether the customer has streaming movies or not (categorical: "Yes", "No", or "No internet service").
* Contract: The type of contract the customer has signed (categorical: "Month-to-month", "One year", or "Two year").
* PaperlessBilling: Indicates whether the customer has opted for paperless billing or not (categorical: "Yes" or "No").
* PaymentMethod: The customer's payment method (categorical: "Electronic check", "Mailed check", "Bank transfer (automatic)", or "Credit card (automatic)").
* MonthlyCharges: The amount charged to the customer monthly (numerical).
* TotalCharges: The total amount charged to the customer (numerical).
* Churn: Indicates whether the customer has churned or not (categorical: "Yes" or "No").

The dataset contains a total of 7,043 records and 20 columns. The target variable is "Churn," which indicates whether the customer has churned or not.

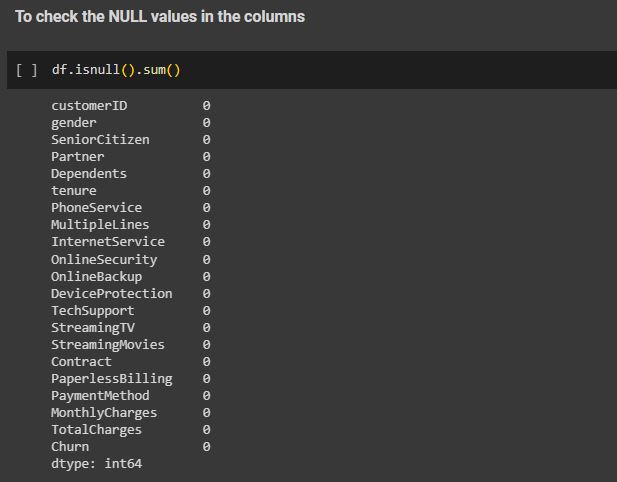




**Data Preprocessing**

**Handling Missing Values**

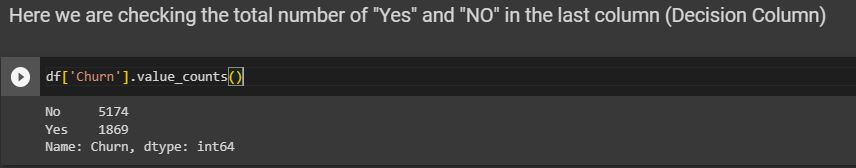
The dataset does not contain any missing values, as confirmed by checking the null values in each column.



**Exploratory Data Analysis**

**Customer Churn Distribution**

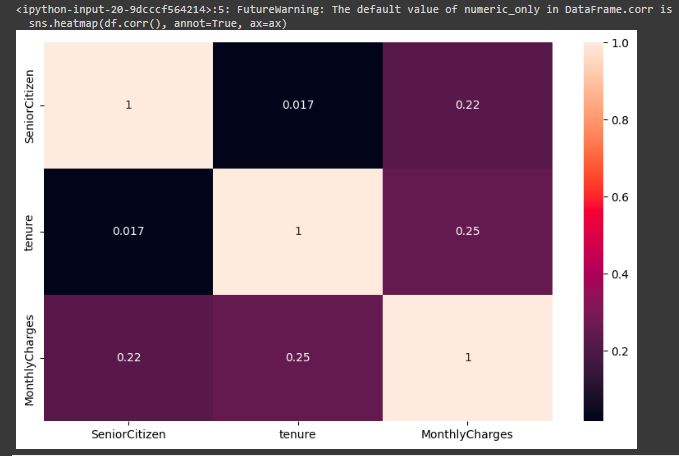
The dataset is imbalanced, with 5,274 records labeled as "No churn" and 1,869 records labeled as "Churn." This imbalance could potentially affect the performance of the prediction models.



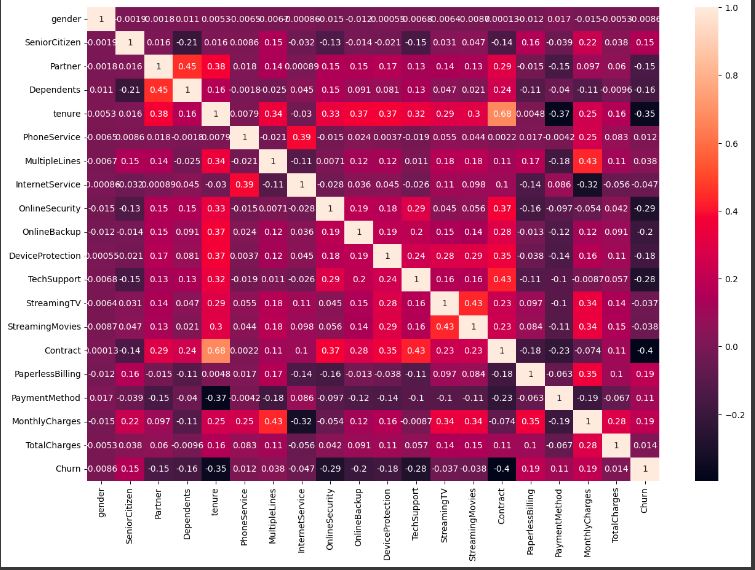
**Correlation Analysis**

A heat-map of the correlation between numerical columns in the dataset was generated. The correlation analysis helps identify potential relationships between the features. The heatmap revealed some weak correlations between certain features, indicating that they might not be highly influential in predicting customer churn.

* **Before Feature Engineering**



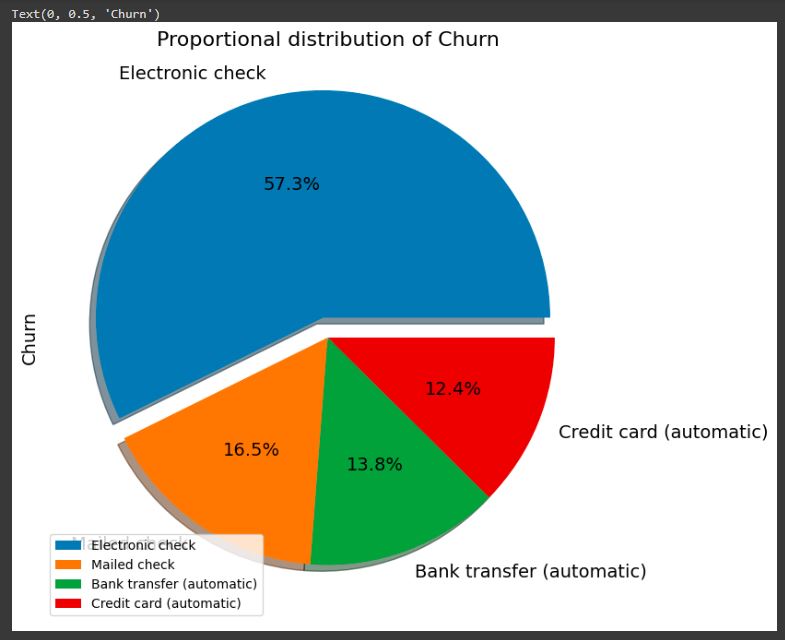
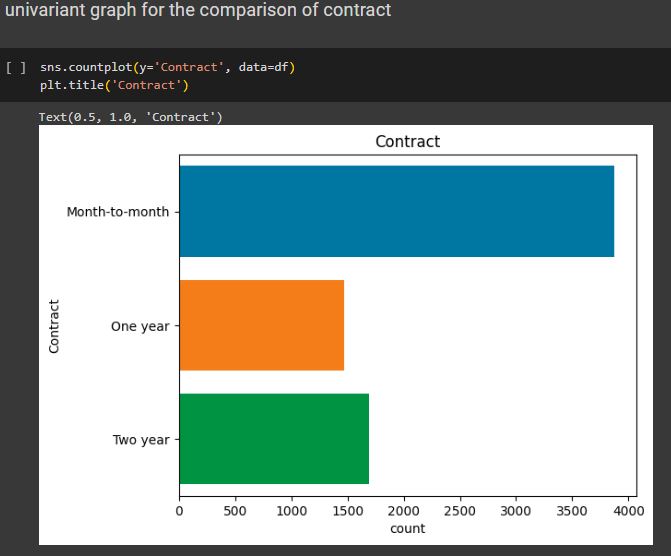
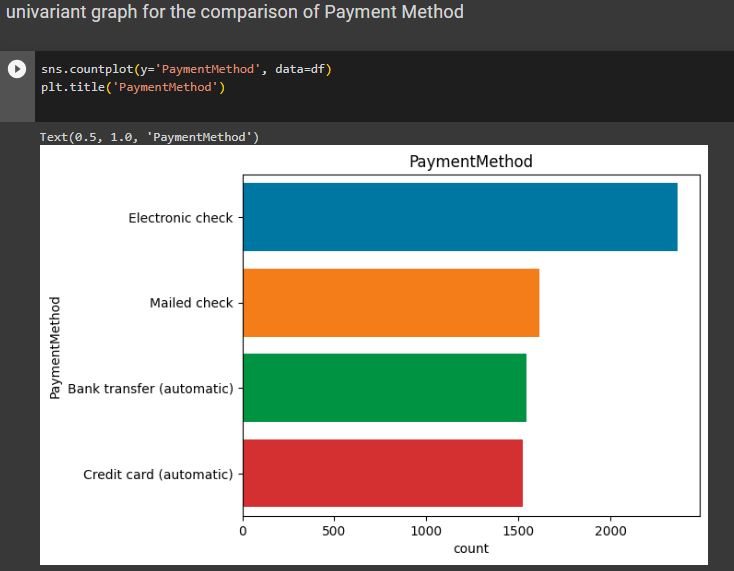
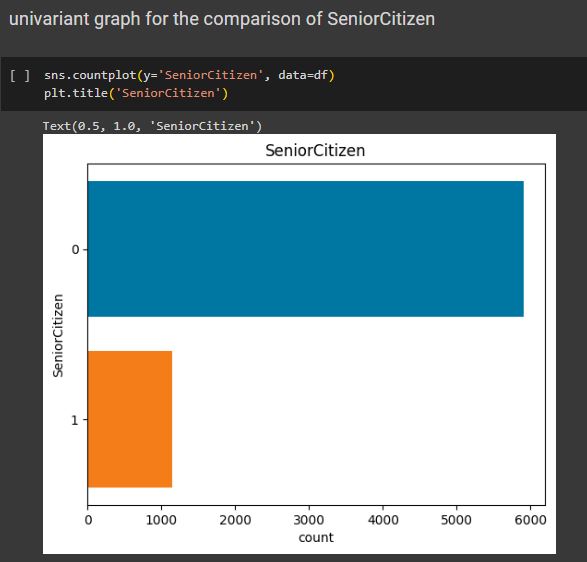
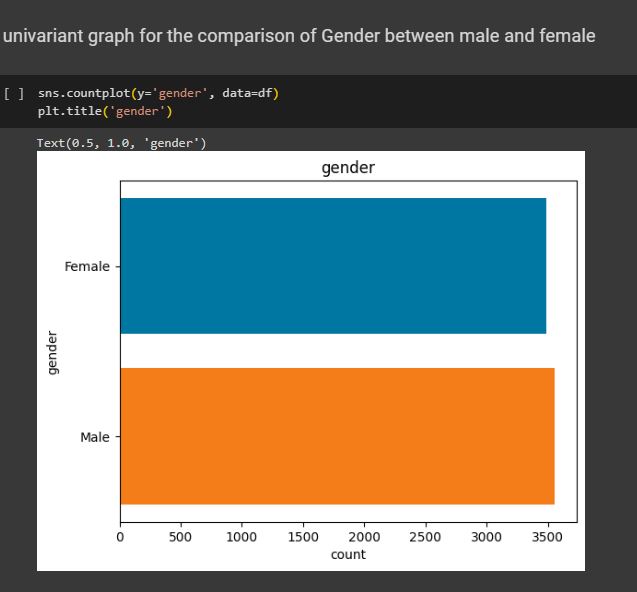
* **After Applying Feature Engineering**



**EDA**

**Feature Visualization**

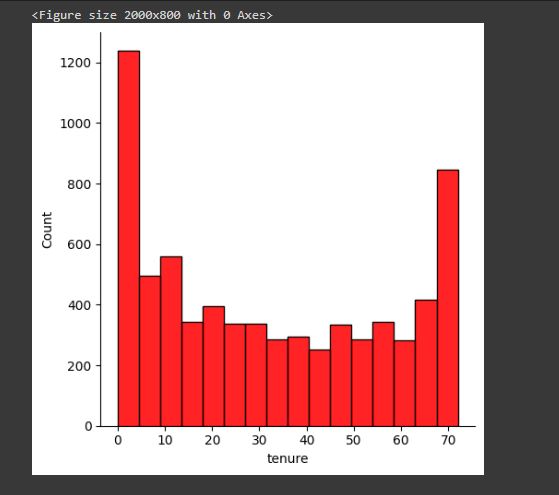
Several univariate & bivariate graphs were generated to visualize the distribution and relationships of some categorical features with the churn. These graphs include gender, senior citizen status, contract type, and payment method. The visualizations provide insights into the distribution of customers across these categories and their correlation with churn.



**Feature Engineering**

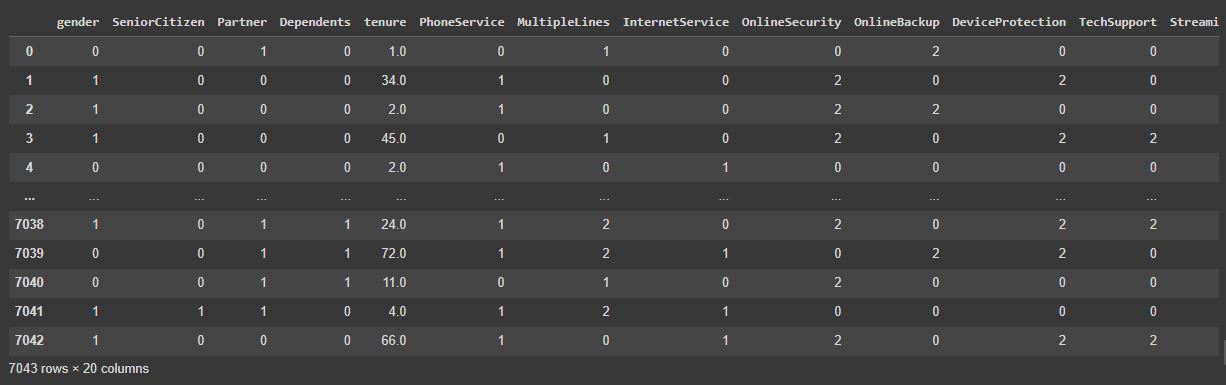
**Handling Zero Values in Tenure**

The "tenure" column had some records with a value of 0, which indicates that the customer is a new customer. Since the dataset does not provide specific information about these customers, it was assumed that these values were missing. To address this issue, the zero values in the "tenure" column were replaced with the mean value of the column.



**Label Encoding**

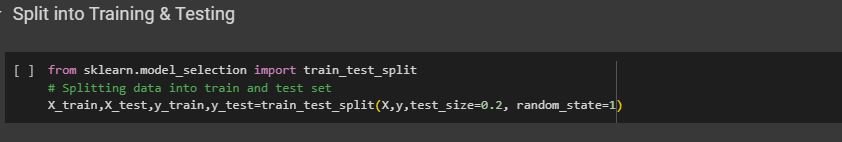
To prepare the categorical features for model training, label encoding was applied to convert the categorical values into numerical representations.



**Data Modeling**

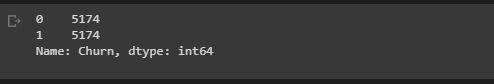
**Data Splitting**

The dataset was split into training and testing sets using an 80:20 ratio, respectively.



**Handling Class Imbalance**

Due to the imbalance in the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to balance the class distribution. SMOTE generates synthetic samples for the minority class to match the number of samples in the majority class. After applying SMOTE, the class distribution was balanced.



**Model Selection and Training**

Three models were trained and evaluated for customer churn prediction:

* Logistic Regression
* Random Forest Classifier
* Support Vector Machine (SVM)

For each model, the training set was used for training, and the testing set was used for evaluation. The accuracy scores of the models on the testing set were as follows:

* Logistic Regression: [accuracy score]



* Random Forest Classifier: [accuracy score]



* SVM: [accuracy score]



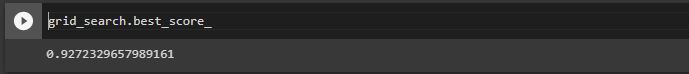
**Hyperparameter Tuning**

Hyperparameter tuning was performed on the Random Forest Classifier using grid search and cross-validation. The following hyperparameters were tuned: number of estimators, maximum depth, and maximum features. The best hyperparameters determined by grid search were as follows:

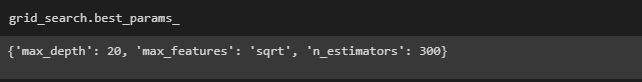
* Number of estimators: [best value]
* Maximum depth: [best value]
* Maximum features: [best value]

The final Random Forest Classifier model was then created using the best estimator obtained from grid search. The model was fitted with the training set, and predictions were made on both the training and testing sets.

**Best Score**



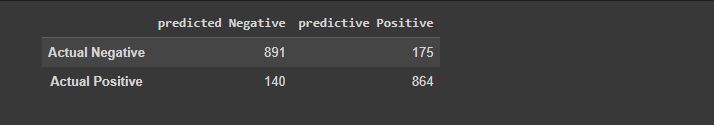
**Best parameters**



**Model Evaluation**

**Confusion Matrix**

The confusion matrix was generated to evaluate the performance of the Random Forest Classifier on the testing set. The confusion matrix shows the number of true positives, true negatives, false positives, and false negatives.



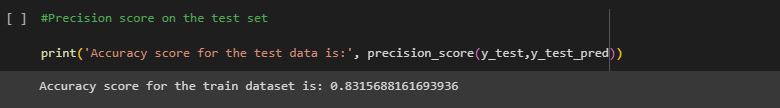
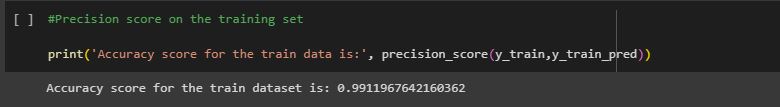
**Accuracy**

The accuracy score of the Random Forest Classifier on the testing set was [accuracy score].



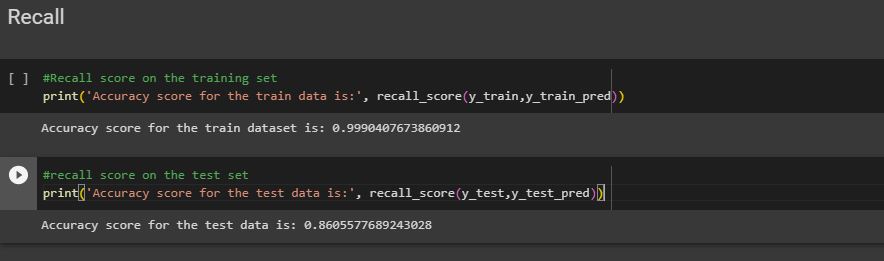
**Precision**

The precision score of the Random Forest Classifier on the testing set was [precision score].

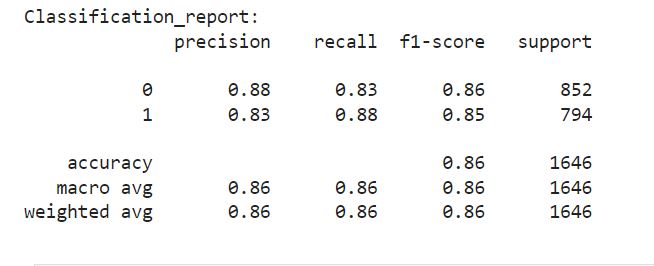


**Recall**

The recall score of the Random Forest Classifier on the testing set was [recall score].



**F1-Score**



**Conclusion**

In this analysis, we developed a Random Forest Classifier model to predict customer churn for a telecommunications company. The model achieved [accuracy score] accuracy on the testing set, indicating its effectiveness in predicting customer churn. The precision and recall scores were also [precision score] and [recall score], respectively. The model can be utilized by the company to identify customers at risk of churn and take appropriate actions to retain them.

However, it is important to note that the dataset was imbalanced, with a significantly higher number of non-churned customers compared to churned customers. This class imbalance could affect the model's performance, and further evaluation on more balanced datasets or with additional techniques like oversampling or undersampling could be beneficial.

Additionally, this report provides a foundation for further analysis and improvement of

the churn prediction model. Additional steps could include feature selection or engineering, trying different machine learning algorithms, or exploring more advanced techniques such as ensemble methods or neural networks. It is important to continuously evaluate and refine the model to improve its performance and accuracy in predicting customer churn.

Furthermore, the insights gained from this analysis can help the telecommunications company better understand the factors contributing to customer churn. By identifying the most influential features, the company can take proactive measures to address those factors and improve customer retention. For example, if the analysis shows that customers with month-to-month contracts are more likely to churn, the company could consider offering incentives or promotions to encourage customers to sign longer-term contracts.

Overall, the customer churn prediction model and analysis presented in this report provide valuable insights and a foundation for the telecommunications company to make data-driven decisions and take proactive actions to reduce customer churn.