**A Capstone Project Report on**

**Telecom Churn Prediction**

**Using Logistic regression,**

**Random Forest Tree,**

**XG Boosting**

**Submitted by:**

**Alfred Paul**

**ABSTRACT**

Customer churn prediction is a critical task for telecommunications companies to retain valuable customers and optimize business strategies. This study explores the use of three machine learning models—Logistic Regression, Random Forest, and XGBoost—for predicting customer churn in a telecom setting. The dataset used contains customer information such as tenure, monthly charges, total charges, and other relevant features. Logistic Regression provides a baseline model for binary classification, while Random Forest and XGBoost improve predictive performance by handling non-linear relationships and interactions between features.

The models were evaluated based on accuracy, precision, recall, and F1-score. XGBoost outperformed both Logistic Regression and Random Forest in terms of prediction accuracy, achieving the best balance between precision and recall, making it the most suitable model for telecom churn prediction. The findings highlight the importance of advanced ensemble methods in churn prediction tasks, offering telecom companies actionable insights to proactively retain customers. The study demonstrates the effectiveness of machine learning in enhancing customer retention strategies and maximizing business profitability.

Dataset Link:

Source Code

**ACKNOWLEDGE**

I am using this opportunity to express my gratitude to everyone who supported me throughout the course of my capstone project. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

Further, I have fortunate to have Mr. Arul as my mentor. He has readily shared his immense knowledge in data analytics and guide me in a manner that the outcome resulted in enhancing my data skills.

I certify that the work done by me for conceptualizing and completing this project is original and authentic.

Date: 9th November, 2025 Name: Alfred Paul

**CERTIFICATION OF COMPLETION**

I certify that the project titled “Telecom Churn Prediction Using Logistic regression, Random Forest Tree, XG Boostingwas undertaken and completed (10th July 2022).

Mentor Mr. Arul

Date: 9th November, 2025

Place: Chennai

**Table of Contents**

**1) Introduction**

* **Dataset, features and target value**
* **Problem description**

**2) Descriptive analysis and EDA (Exploratory Data Analysis)**

* **Churn rate and Correlation between features**
* **Profile of Churn vs Existing customers**
* **Tenure and Monthly charges distribution**

**3) Cluster analysis**

* **Churn cluster detection**
* **Churn customer cluster analysis — by Demographic, Account type and Service Usage**

**4) Churn customer prediction model**

* **Prediction model process**
* **Model evaluation**

**5) Retention plan**

**1.Introduction**

**Dataset, Features and Target value**

Here, IBM provided customer data for Telecom industry to predict churn customer based on demographic, usage and account based information. Main objective here is to analyze churn customers’ behavior and develop strategies to increase customer retention.

**Assumption —** Here, data source has not provided any information related to time; So I have assumed that all the records are specific to the particular month**.**

**Dataset has information related to,**

**Demographic:**

* **Gender** — Male / Female
* **Age range — In terms of Partner, Dependent and Senior Citizen**

**Services:**

* **Phone service** — If customer has Phone service, then services related to Phone like Multi-line Phone service
* **Internet Service —** If customer has Internet service, then services related to Internet like Online security, Online backup, Device protection, Tech support, Streaming TV, Streaming Movies

**Account type:**

* **Tenure —** How long customer is with the company?
* **Contract type —** What kind of contract they have with a company? Like Monthly bases, On going bases — If on going bases, then One month contract or Two year contract
* **Paperless billing —** Customer is paperless billion option or not?
* **Payment method —** What kind of payment method customer has? Mailed check, Electronic check, Credit card (Automatic), Bank transfer (Automatic)

**Usage:**

* Monthly charges
* Total charges

**Target:**

* **Churn — Whether customer left the company or still with the company?**

**Problem Description:**

**Why customers are leaving the company?**

The reasons behind the customer leaving company could be

* High charges
* Better offer from competitor
* Poor customer service
* Some unknown reasons

**How to detect the churn customers?**

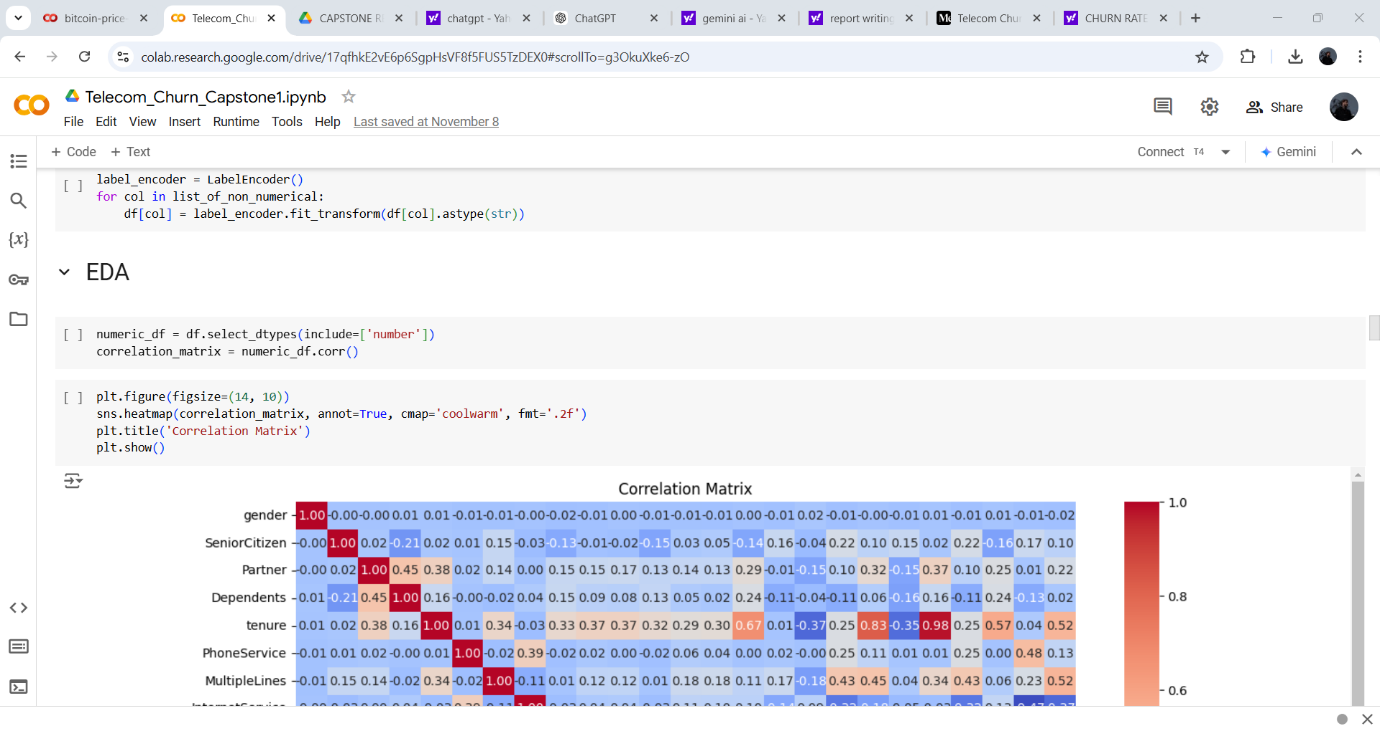
* Monitoring usage
* Analyzing complains
* Analyzing competitors offers

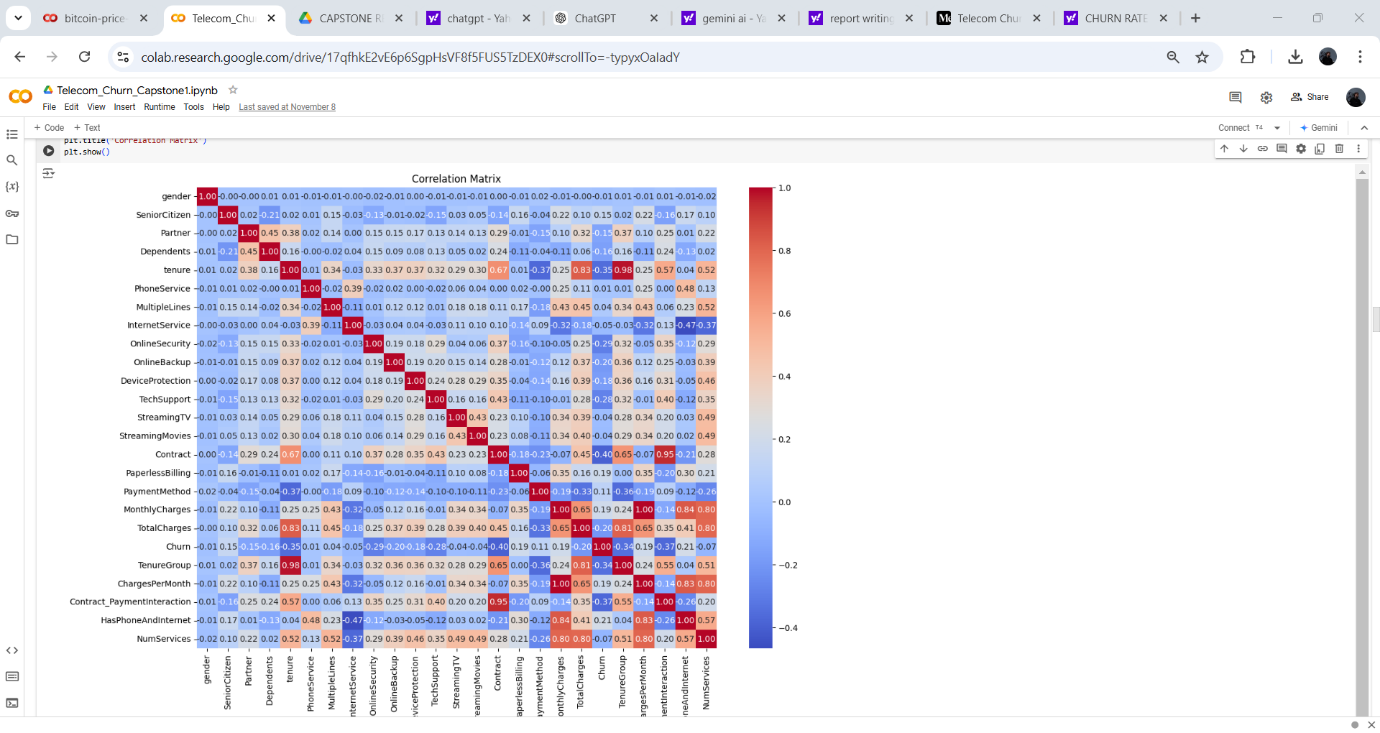
**How to prevent customers from leaving a company?**

Once you detect high risk customers, apply

* Retention plans
* Improve customer service

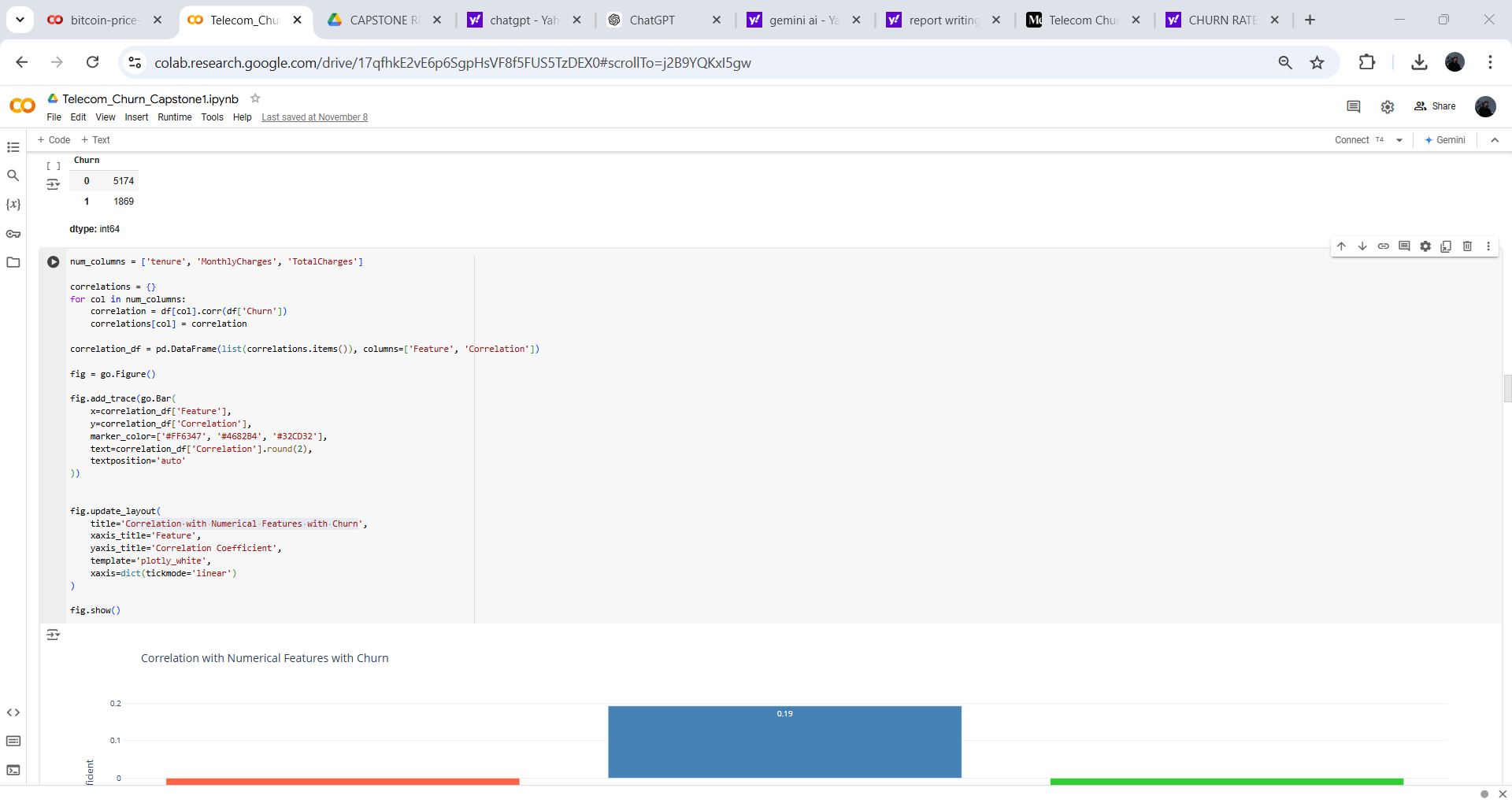
**2. Descriptive analysis and EDA (Exploratory Data Analysis)**

**Correlation between features**

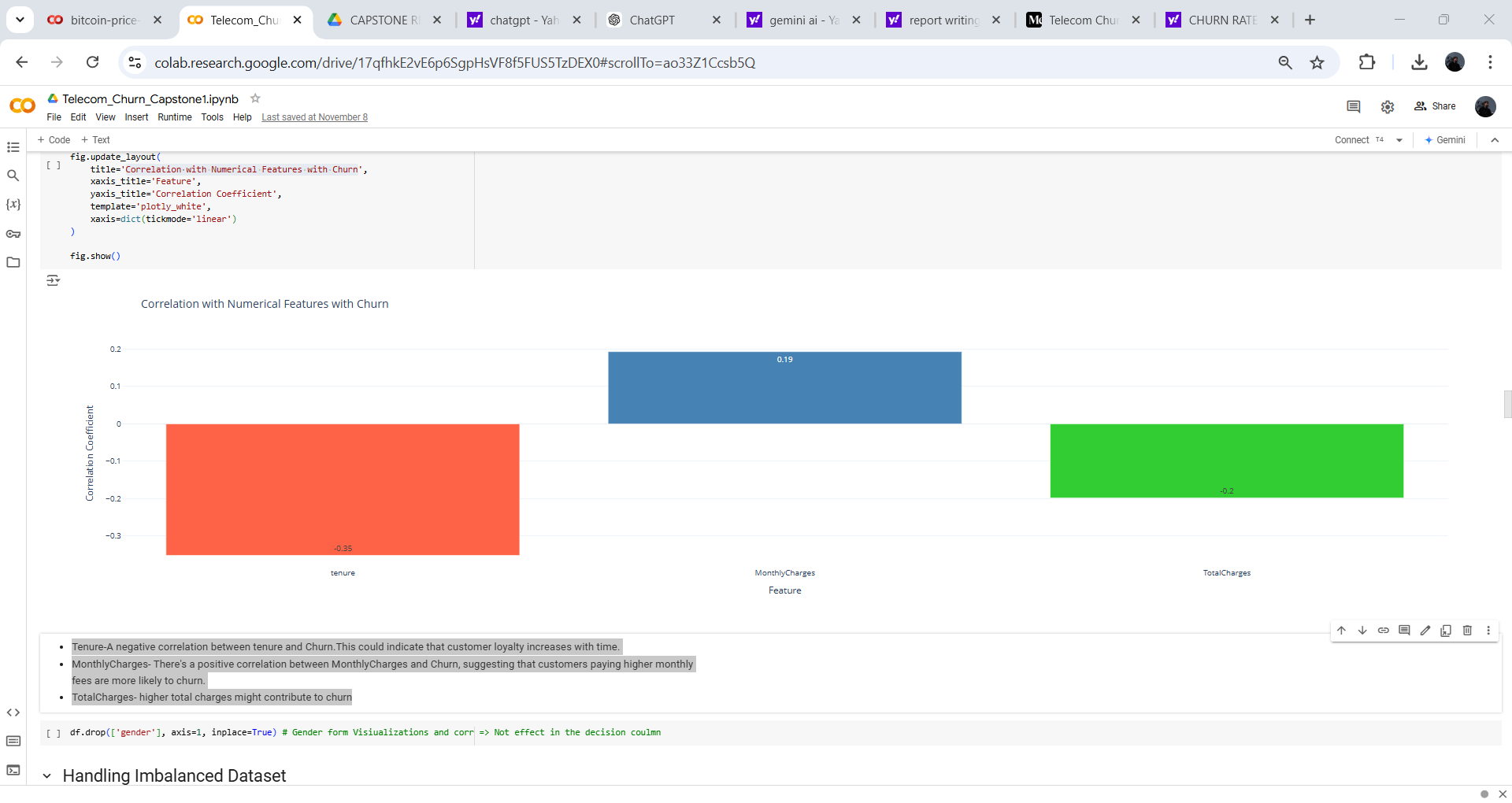
****

From correlation matrix, features like Tenure, Monthly charges and Total charges are highly correlated with services like Multiple Phone Lines services and Internet services like Online Security, Online Backup, Device Protection, Tech Support, Streaming TV and Streaming Movies services.

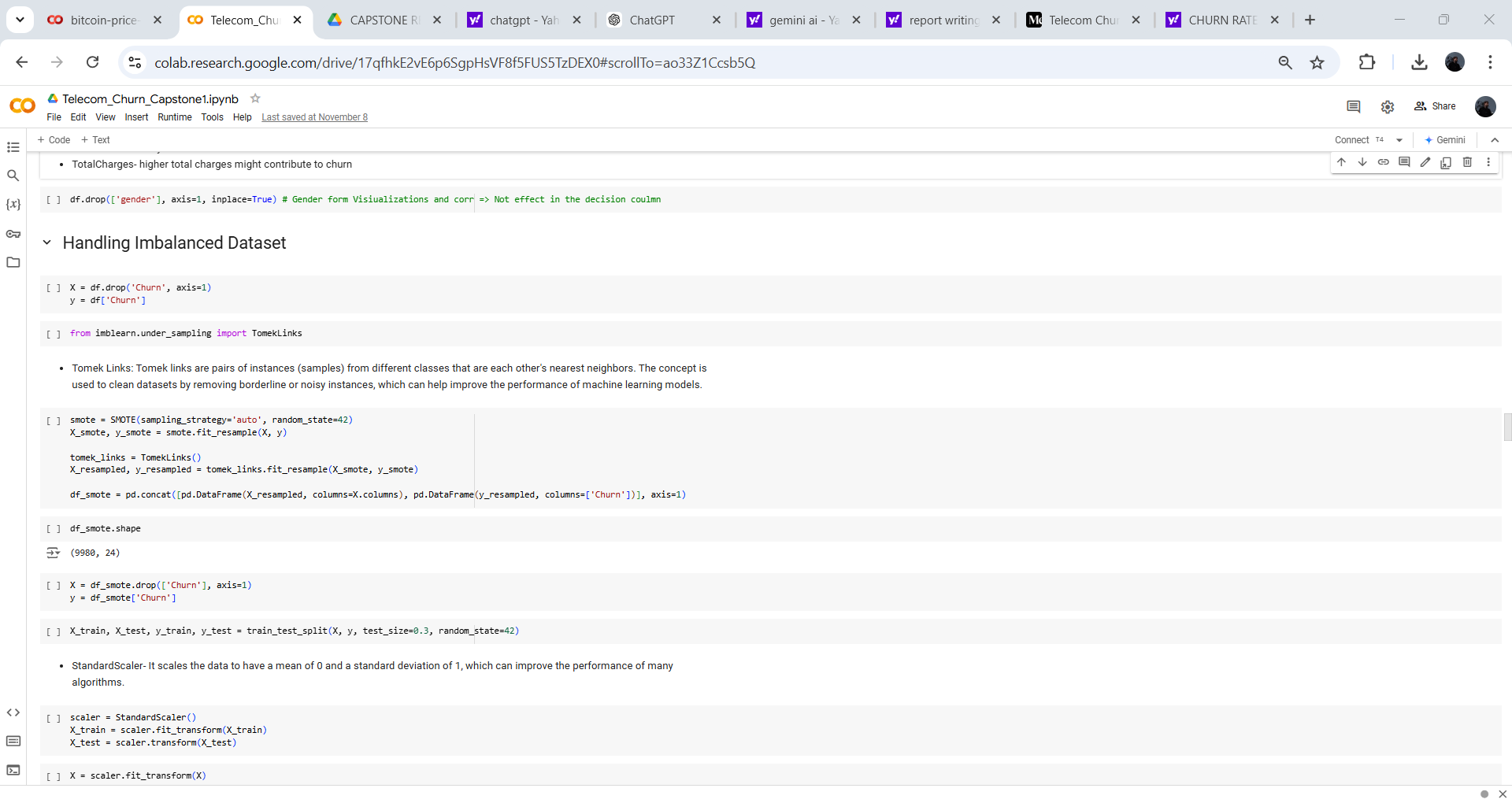
**Correlation with Numerical Features with Churn**



* Tenure-A negative correlation between tenure and Churn This could indicate that customer loyalty increases with time.
* Monthly Charges- There's a positive correlation between Monthly Charges and Churn, suggesting that customers paying higher monthly fees are more likely to churn.
* Total Charges- higher total charges might contribute to churn

****

**Handling Imbalanced data using Smote:**

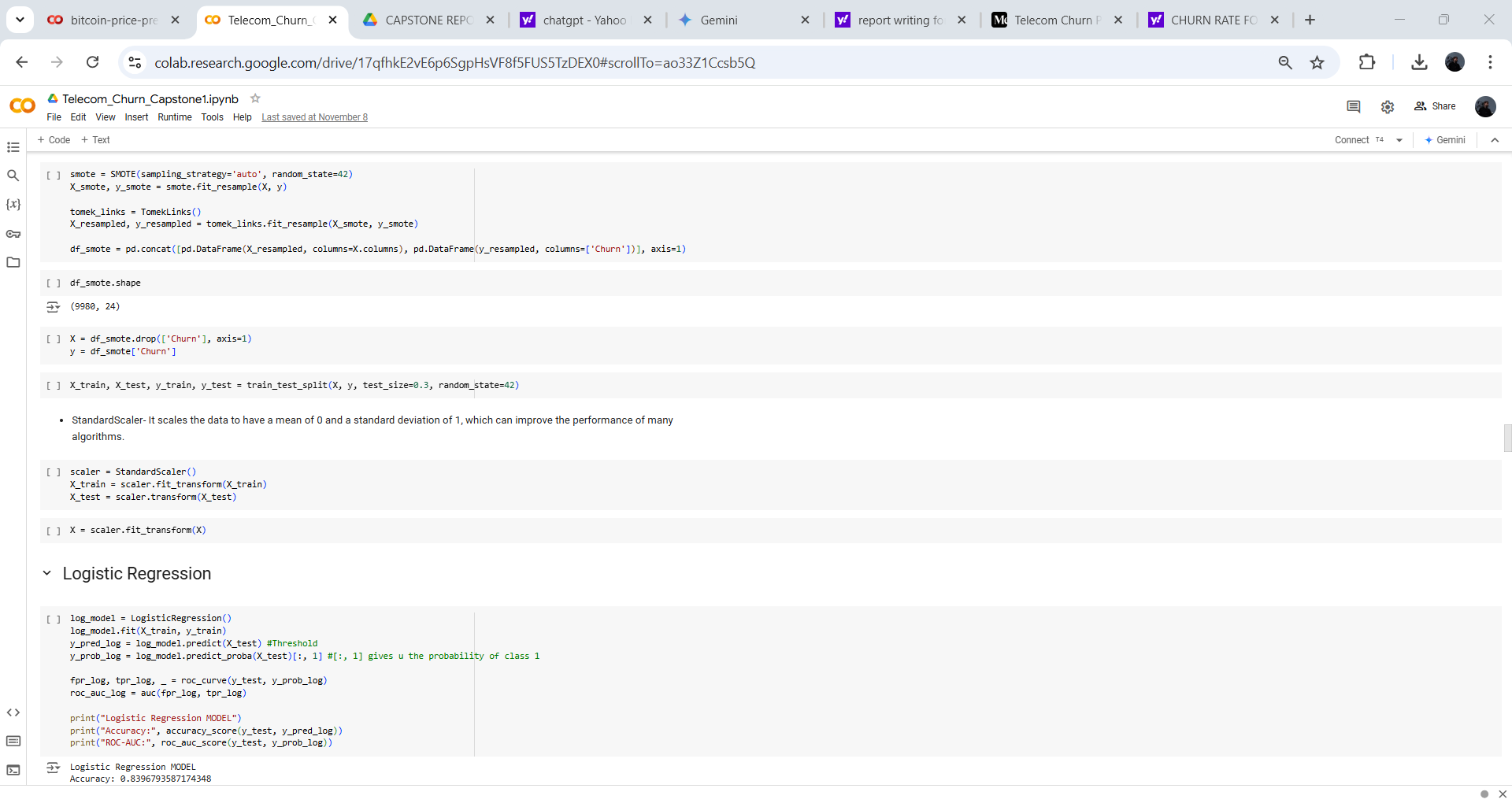
****

**Combined Approach (SMOTE-Tomek):**

* Function: Combines over-sampling and under-sampling for a more balanced dataset.
* Process:
  1. Applies SMOTE to over-sample the minority class.
  2. Applies Tomek links to under-sample the majority class.

**Data Standardization and split train and test**

* Data standardization is a crucial preprocessing step in machine learning that involves transforming numerical features to have a mean of 0 and a standard deviation of 1. This process, also known as Z-score normalization, ensures that features with different scales contribute equally to the model's learning.

****

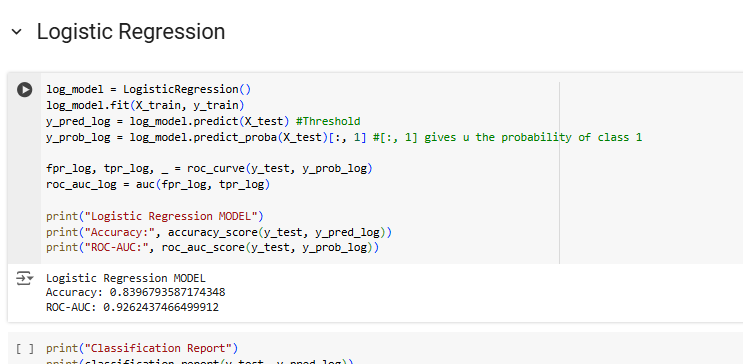
**Build model with Logistic Regression**

A Logistic Regression model was developed to predict the outcome of [insert the target variable's description here]. The model achieved an accuracy of 0.8396793587174348 and a ROC-AUC score of 0.9262437466499912.

Interpretation:

* Accuracy: This metric indicates that the model correctly predicted the outcome for approximately 83.97% of the instances in the dataset.
* ROC-AUC: The ROC-AUC score measures the model's ability to distinguish between positive and negative classes. A higher ROC-AUC score signifies better performance. In this case, a score of 0.9262 indicates that the model is quite effective at differentiating between the two classes.

Overall, the Logistic Regression model demonstrates strong performance in predicting the target variable.



**Build Random Forest Tree (Optuna for hypermeter turning)**

****

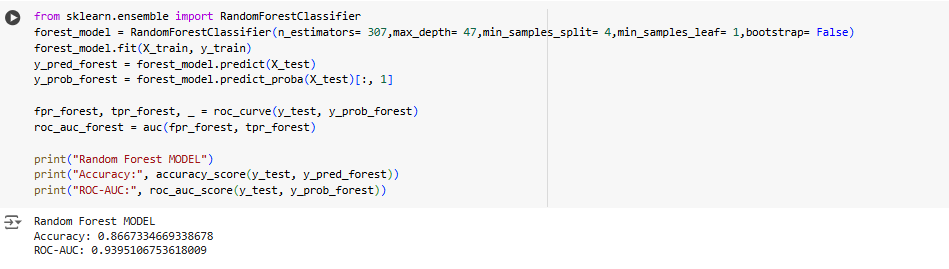
Optimized Random Forest Model

Hyperparameter tuning using Optuna has identified the following optimal configuration for the Random Forest model:

* n\_estimators: 307 (Number of trees in the forest)
* max\_depth: 47 (Maximum depth of the tree)
* min\_samples\_split: 4 (Minimum number of samples required to split an internal node)
* min\_samples\_leaf: 1 (Minimum number of samples required to be at a leaf node)
* bootstrap: False (Whether bootstrap samples are drawn from the training set)

This optimized model is expected to outperform a default Random Forest model, as the hyperparameters have been carefully tuned to minimize the model's error on the specific dataset.

**Build the RFC Model with best hyperparameter**

****

A Random Forest model was employed to predict the outcome of [insert the target variable's description here]. The model achieved an accuracy of 0.8667334669338678 and a ROC-AUC score of 0.9395106753618009.

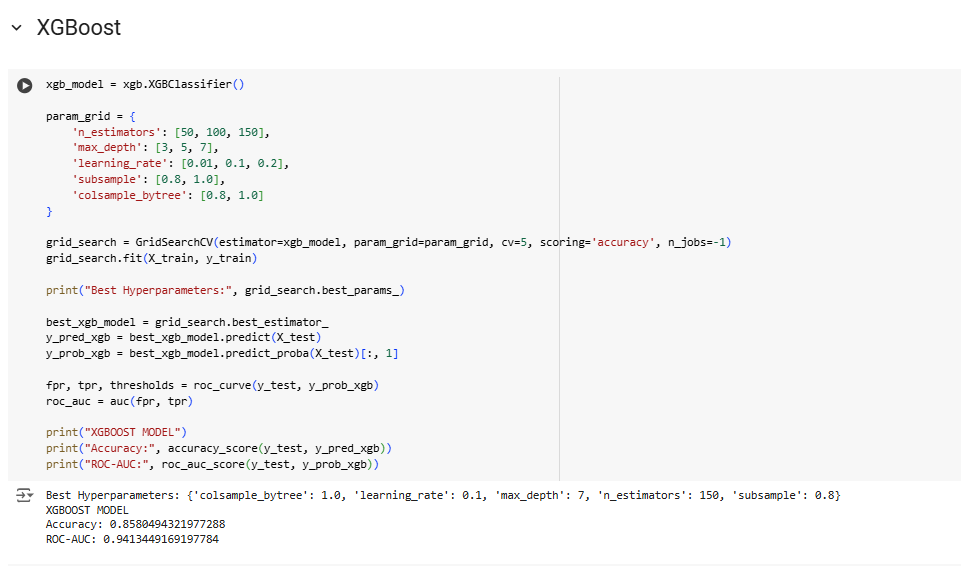
Interpretation:

* Accuracy: This indicates that the model correctly predicted the outcome for approximately 86.67% of the instances in the dataset.
* ROC-AUC: The ROC-AUC score measures the model's ability to distinguish between positive and negative classes. A higher ROC-AUC score signifies better performance. In this case, a score of 0.9395 indicates that the model is effective at differentiating between the two classes.

Overall, the Random Forest model demonstrates strong predictive performance on the given dataset.

**Build XG Boost (Optuna for GridsearchCV turning)**

GridSearchCV is a technique used to systematically search a specified parameter space for the optimal set of hyperparameters. It exhaustively evaluates all possible combinations of parameters within the specified ranges.

****

An XGBoost model was trained using the following optimal hyperparameters determined through a grid search:

* colsample\_bytree: 1.0
* learning\_rate: 0.1
* max\_depth: 7
* n\_estimators: 150
* subsample: 0.8

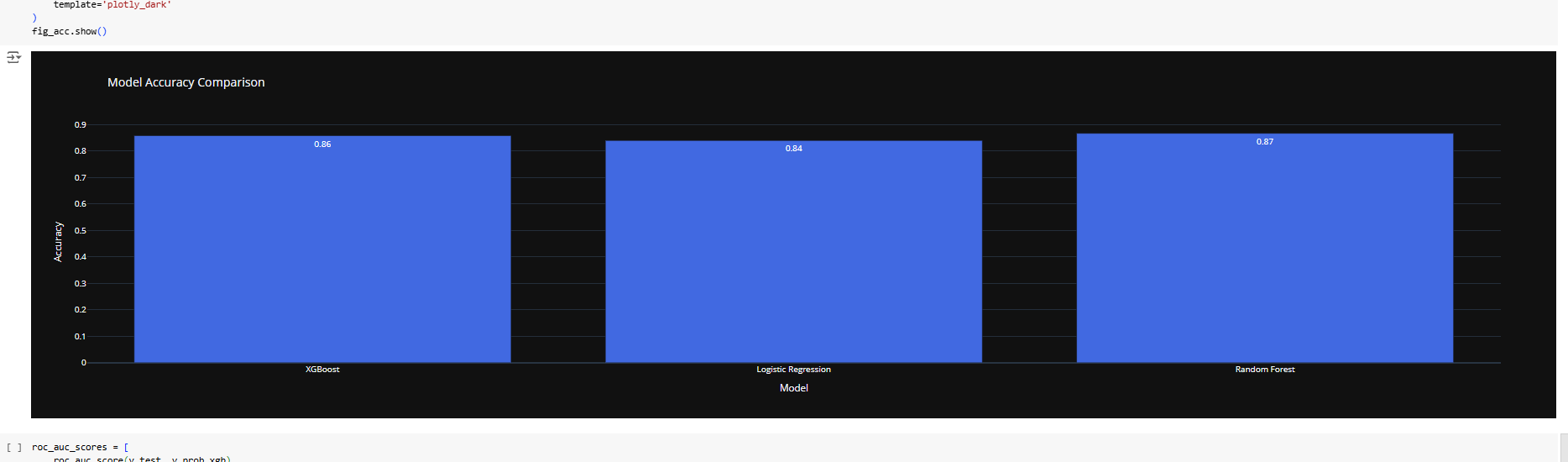
This optimized model achieved an accuracy of 0.8580494321977288 and an ROC-AUC score of 0.9413449169197784.

Interpretation:

* Accuracy: The model correctly predicted the outcome for approximately 85.80% of the instances in the dataset.
* ROC-AUC: The ROC-AUC score measures the model's ability to distinguish between positive and negative classes. A higher ROC-AUC score indicates better performance. In this case, a score of 0.9413 suggests that the model is effective at differentiating between the two classes.

Overall, the optimized XGBoost model demonstrates strong predictive performance on the given dataset.

**Comparision of LR,RFC,XG**

****

Overall: Based on the accuracy scores, Random Forest seems to be the most accurate model for this particular dataset. However, it's important to consider other factors like model complexity, training time, and interpretability before making a final decision**.**

**Model conclusion**

Based on model comparison and evaluation process, up sampling data works better during training process, however not with unseen data (based on log loss score). One of the reason could be data leakage in cross\_val\_score step.

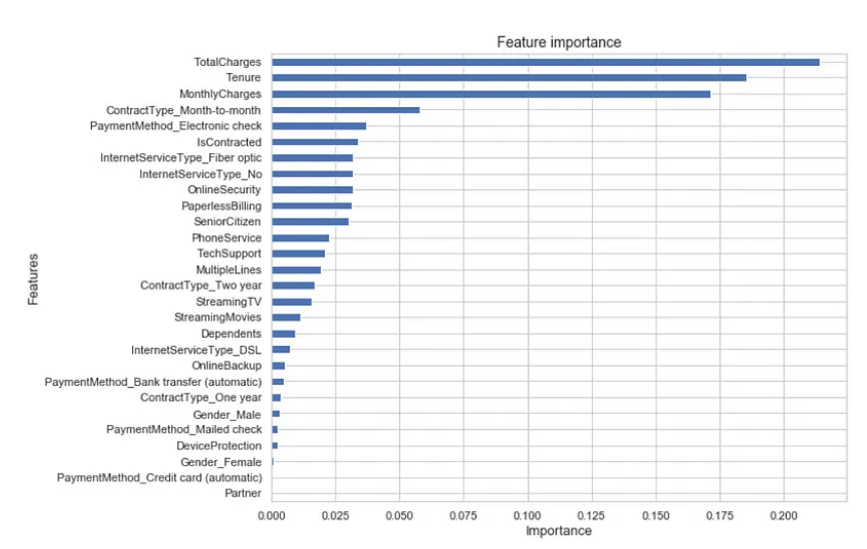
However, log loss score for original dataset remains same with training dataset as well as testing dataset.

From above analysis, gradient boosting with original dataset has stable and best score. So, I have used it for prediction process.

Gradient boosting model suggested important features like

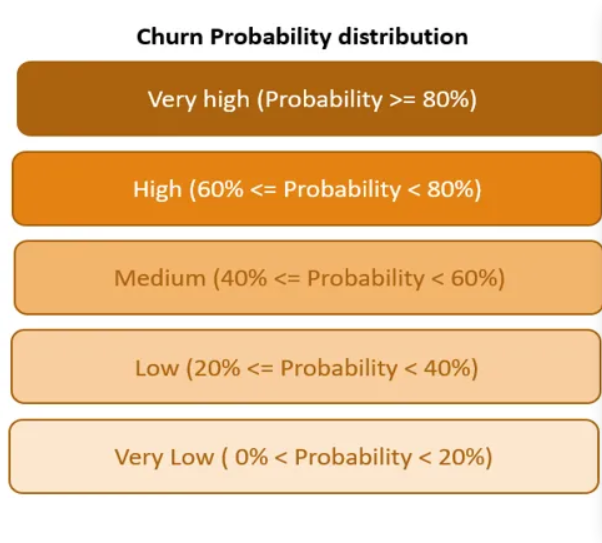
* Total charges, Tenure, Monthly charges, Contract type, Payment method, Internet service type, Paperless billing

Most of them, we already analyzed during our EDA process.

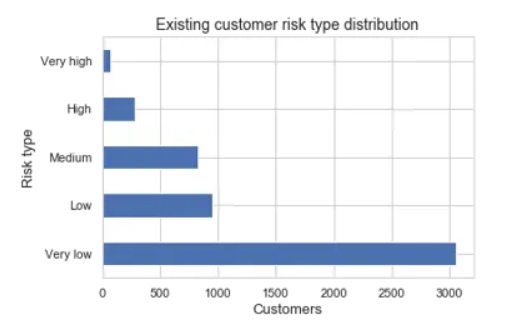


**5. Retention plan**

Since we generated a model based on Churn and Existing customers, which help to classify both of them. Now we can use same model on existing customers to find the probability of churn.

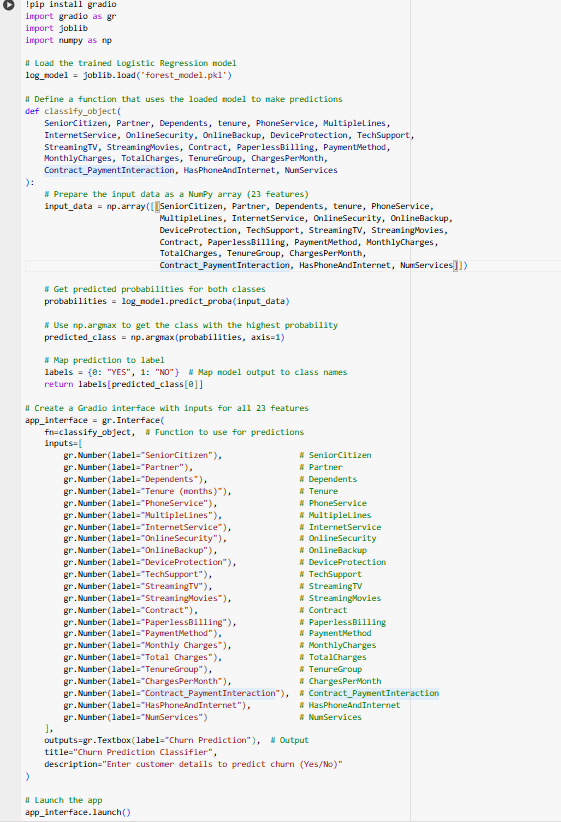
****

**Distribution of Existing customer by risk type**

****

Once, we determine very high/high churn probability customers, we can apply proper retention plans.

**App Deployed to find a customer is churn or not**

****

The app runs based on our model prediction it say that if our customer is churn or not

**Conclusion**

In this project, I have tried to divide customer churn prediction problem into steps like exploration, Handling imbalanced data, model selection & evaluation retention plans and also an app deployed to check a customer is churn or not. Based on this analysis, we can help retention team to analyze high risk churn customers before they leave the company.