Project Report

**TELECOM CHURN PREDICTION MODEL USING MACHINE LEARNING**

## A PROJECT REPORT

***Submitted by***

***Tushar Narula(24bda70353) Rajarshi das(24bda70351) Ayush Upadhyay(24bda70346) Aniket das (24bda70336) Chirag sharma(24bda70332)***

***in partial fulfillment for the award of the degree of***

# BACHELOR OF ENGINEERING

**IN**

COMPUTER SCIENCE ENGINEERING (DATA SCIENCE)

## Chandigarh University

APRIL 2025



# BONAFIDE CERTIFICATE

Certified that this project report “ **TELECOM CHURN PREDICTION MODEL USING MACHINE LEARNING** ” is the bonafide work of “ **TUSHAR RAJARSHI AYUSH ANIKET CHIRAG ”** who carried out the project work under my/our supervision.

#### SIGNATURE

Dr.Urvashi Garg

#### HEAD OF THE DEPARTMENT

Computer Science and Engineering

#### SIGNATURE

Mrs.Neeru

#### SUPERVISOR

Assistant Professor

Computer Science and Engineering

Submitted for the project viva-voce examination held on

**INTERNAL EXAMINER EXTERNAL EXAMINER**

# TABLE OF CONTENTS

List of Figures 6

List of Tables 7

Abstract …………………………………………………………………………………………….. 8

Graphical Abstract ………………………………………………………………………………….. 9

Abbreviations ……………………………………………………………………………………… 10

Symbols …………………………………………………………………………………………… 11

CHAPTER 1. INTRODUCTION 12

* 1. [Identification of Client/ Need/ Relevant Contemporary issue 1](#_TOC_250017)2
  2. [Identification of Problem 1](#_TOC_250016)2
  3. [Identification of Tasks 1](#_TOC_250015)2
  4. [Timeline 1](#_TOC_250014)2
  5. [Organization of the Report 1](#_TOC_250013)2

CHAPTER 2. LITERATURE REVIEW/BACKGROUND STUDY 13

* 1. [Timeline of the reported problem 13](#_TOC_250012)
  2. [Existing solutions 1](#_TOC_250011)3
  3. [Bibliometric analysis 1](#_TOC_250010)3
  4. [Review Summary 13](#_TOC_250009)
  5. [Problem Definition 1](#_TOC_250008)3
  6. [Goals/Objectives 1](#_TOC_250007)3

CHAPTER 3. DESIGN FLOW/PROCESS 14-15

* 1. [Evaluation & Selection of Specifications/Features 13](#_TOC_250006)
  2. [Design Constraints 13](#_TOC_250005)
  3. [Analysis of Features and finalization subject to constraints 13](#_TOC_250004)
  4. [Design Flow 14](#_TOC_250003)
  5. [Design selection 14-1](#_TOC_250002)5
  6. [Implementation plan/methodology 1](#_TOC_250001)5

CHAPTER 4. RESULTS ANALYSIS AND VALIDATION 16

* 1. [Implementation of solution 1](#_TOC_250000)6

CHAPTER 5. CONCLUSION AND FUTURE WORK 17-18

* 1. [Conclusion 1](#_bookmark0)7
  2. [Future work 17-18](#_bookmark1)

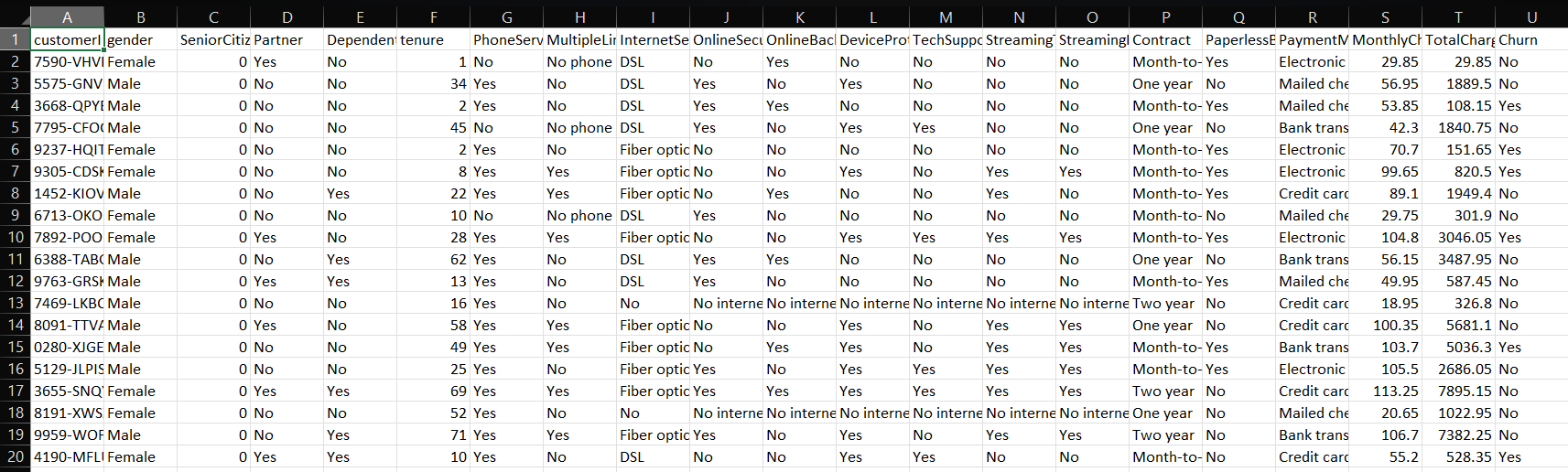
REFERENCES 19

APPENDIX 20

1. Plagiarism Report 20

USER MANUAL 21

|  |  |
| --- | --- |
| Figure 3.1 | Workflow of the Churn Prediction Model  corresponds to the Design Flow process in Chapter 3. |
| Figure 3.2 | Feature Importance Plot (Random Forest / XGBoost)  shows the impact of features like tenure, contract type, monthly charges, etc. |
| Figure 4.1 | Confusion Matrix of the Best Performing Model (XGBoost)  shows the evaluation results of your model using a confusion matrix. |
| Figure 5.2.1 | Plagiarism Report |

****

**.**

**.**

**.**

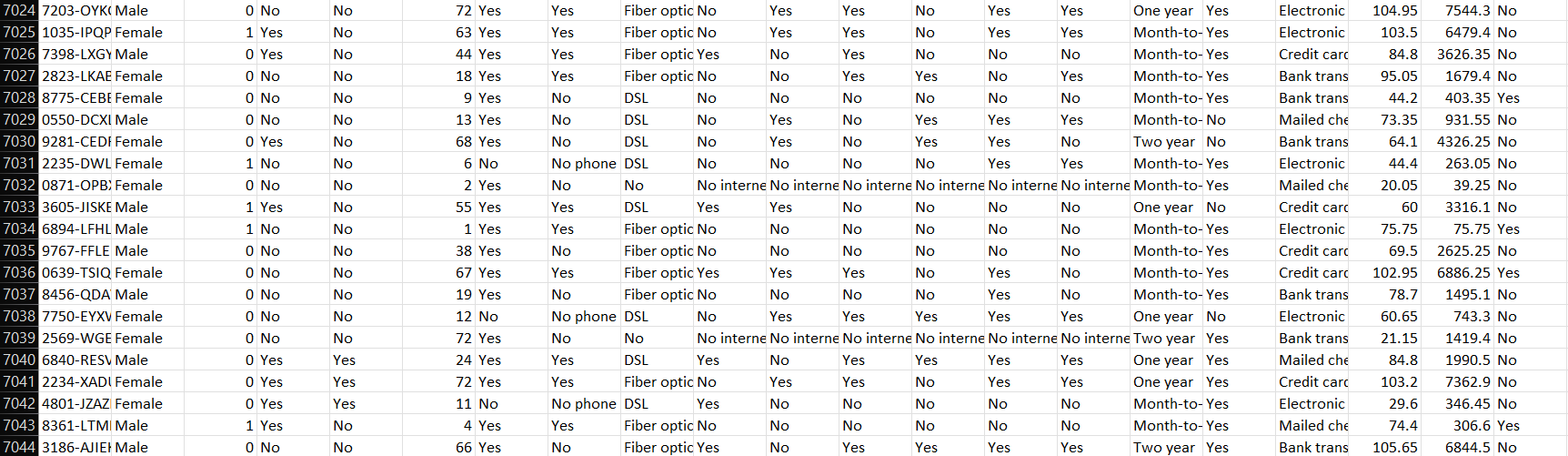
**.**

**.**

**.**

**.**

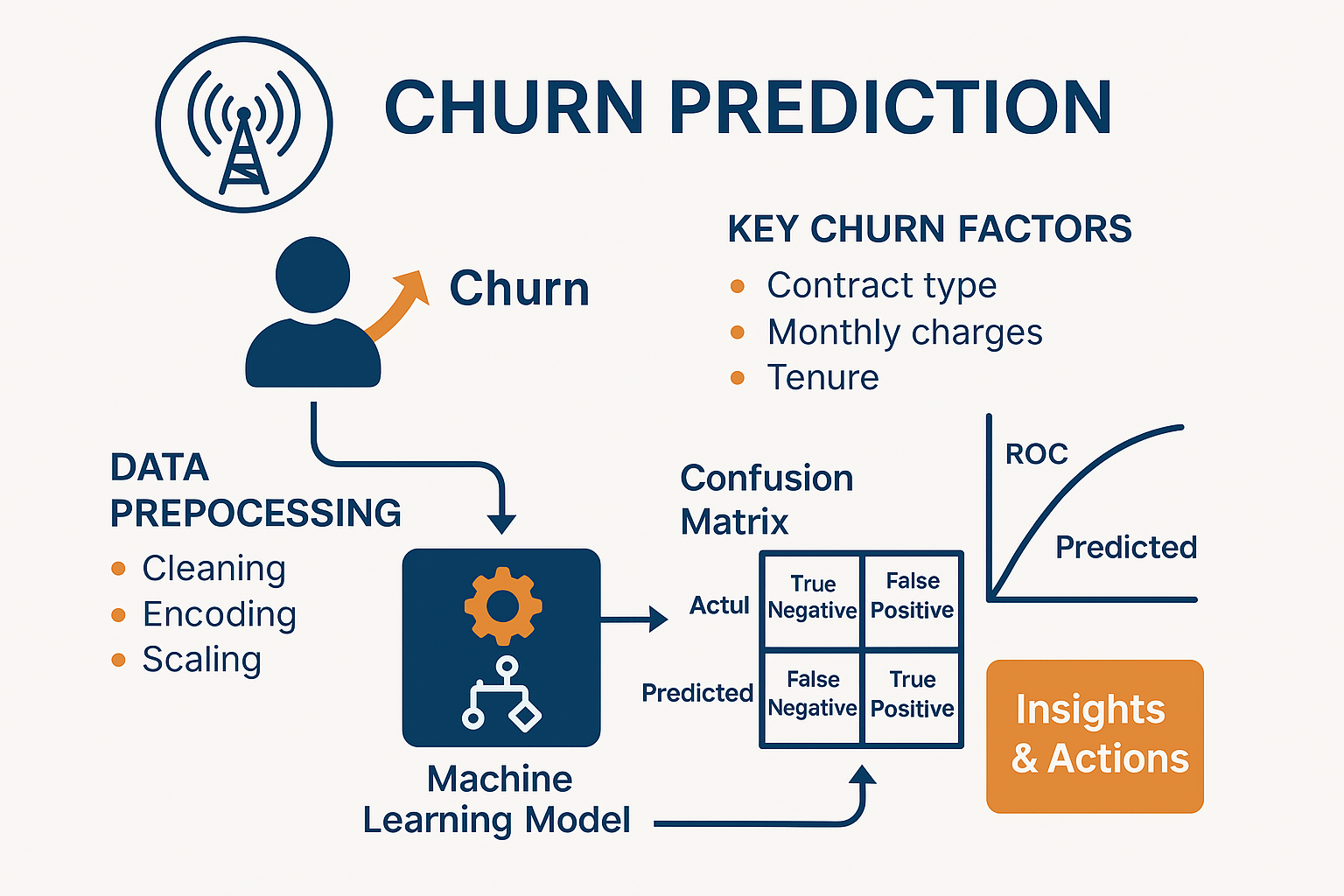
**.**

****

# ABSTRACT

In the competitive telecom sector, retaining customers is essential to maintaining profitability. Customer churn—loss of clients or subscribers—entails high revenue loss and higher customer acquisition prices. This project uses predictive analytics and machine learning to anticipate customer churning risk beforehand. Through Random Forest and XGBoost algorithms applied to actual telecom data, we intend to construct an accurate predictive model. With the comprehensive preprocessing of data and investigation of important factors of churn like contract type, monthly fees, and tenure, the model provides actionable insights. The solution, deployed in Google Colab, helps telecom companies improve retention policies, lower churn rates, and make business decisions.

**GRAPHICAL ABSTRACT**

****

**ABBREVIATIONS**

|  |  |
| --- | --- |
| ML | Machine Learning |
| RF | Random Forest |
| XGBoost | Extreme Gradient Boosting |
| SMOTE | Synthetic Minority Over-sampling Technique |
| ROC | Receiver Operating Characteristic |
| TP | True Positive |
| TN | True Negative |
| FP | False Positive |
| FN | False Negative |
| CV | Cross Validation |
| GridSearchCV | Grid Search with Cross Validation |
| KPI | Key Performance Indicator |
| CSV | Comma-Separated Values |
| GUI | Graphical User Interface |
| API | Application Programming Interface |

**SYMBOLS**

|  |  |
| --- | --- |
| → (Arrow) | Used in flow diagrams and phase transitions (e.g., Data Preprocessing → Modeling) |
| TP | True Positive – churn correctly predicted as churn |
| TN | True Negative – no churn correctly predicted as no churn |
| FP | False Positive – no churn wrongly predicted as churn (Type I Error) |
| FN | False Negative – churn wrongly predicted as no churn (Type II Error) |
| % (Percent) | Used in metrics like accuracy, precision, and recall |
| = | Indicates assignment or equality in metric formulas or feature handling |
| ( ) | Used in model names (e.g., GridSearchCV), formulas, or performance metrics |

# INTRODUCTION

## Identification of Client / Need / Relevant Contemporary Issue

Customer churn is a major challenge for telecom companies. With growing competition, retaining existing users is more cost-effective than acquiring new ones. Machine learning offers a modern solution to predict churn based on customer data, helping companies take preventive actions.

## Identification of Problem

Telecom companies often struggle to identify which customers are likely to leave. Manual methods are inefficient and inaccurate. The goal is to build a machine learning model that can predict churn using customer behavior, usage, and service data.

## Identification of Tasks

* + - Collect and clean telecom customer data
    - Analyze key features linked to churn
    - Train and compare machine learning models
    - Evaluate model performance
    - Provide insights for reducing churn

## Timeline

|  |  |
| --- | --- |
| Week | Task |
| 1-2 | Research and data collection |
| 3-4 | Data cleaning and analysis |
| 5-7 | Model development |
| 8-9 | Evaluation and tuning |
| 10-12 | Report writing and presentation |

## Organization of the Report

* + - Chapter 1: Introduction
    - Chapter 2: Literature Review
    - Chapter 3: Methodology
    - Chapter 4: Implementation and Results
    - Chapter 5: Conclusion and Future Work

# LITERATURE REVIEW/BACKGROUND STUDY

## Timeline of the Reported Problem

The issue of customer churn in telecom has been recognized since the early 2000s, with companies noticing major revenue losses due to customer dissatisfaction and competition. With the rise of big data and machine learning in the 2010s, predictive churn modeling became more effective and widely researched.

## Existing Solutions

Traditional methods involved manual analysis and rule-based systems. Recently, machine learning techniques such as logistic regression, decision trees, random forests, and neural networks have shown better performance in identifying patterns in customer behavior to predict churn accurately.

## Bibliometric Analysis

Research in churn prediction has grown steadily over the last decade. A significant number of studies have been published in journals like IEEE Access, Elsevier's Expert Systems with Applications, and Springer. Most research focuses on using supervised learning models with real- world telecom datasets (e.g., from Kaggle or telecom providers).

## Review Summary

The literature indicates that machine learning models outperform traditional statistical methods in churn prediction. However, challenges such as data imbalance, feature selection, and model interpretability remain areas of active research.

## Problem Definition

Despite advancements, many telecom companies still lack accurate and real-time churn prediction systems. This project aims to build an efficient machine learning model that predicts customer churn based on historical usage and service data.

## Goals / Objectives

* + - Analyze telecom customer data to identify churn indicators
    - Apply and compare multiple ML algorithms
    - Select the best-performing model
    - Provide insights to help telecom companies reduce churn

# DESIGN FLOW/PROCESS

### Evaluation & Selection of Specifications / Features

Feature selection is crucial for accurate predictions. Based on data analysis, features like customer tenure, monthly charges, contract type, payment method, internet service, and customer support calls were selected. These attributes have shown strong correlation with churn behaviour in previous studies.

### Design Constraints

The project faced several constraints:

* + - Data availability and quality: Missing or inconsistent values
    - Imbalanced classes: Fewer churners compared to non-churners
    - Model complexity vs. interpretability: Trade-off between performance and transparency
    - Time & computational resources: Limited training and testing time

### Analysis of Features and Finalization Subject to Constraints

After handling missing values and analysing feature importance using correlation analysis and tree- based methods, the final feature set was chosen. Redundant and less informative features were removed to improve model efficiency while adhering to processing constraints.

### Design Flow

The overall process includes:

1. Data Collection
2. Data Cleaning and Preprocessing
3. Feature Selection and Engineering
4. Model Selection and Training
5. Model Evaluation and Tuning
6. Deployment Planning

### Design Selection

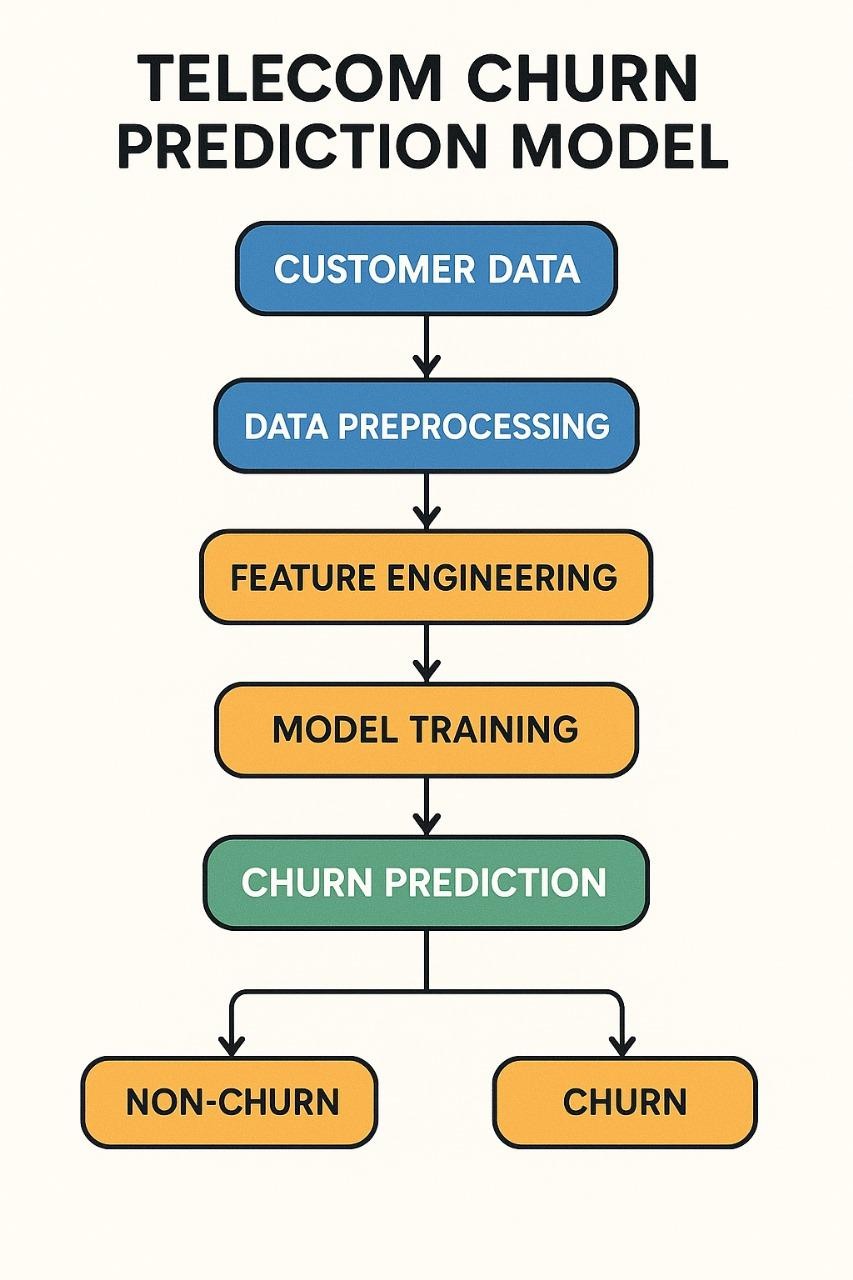
Several machine learning models were tested:

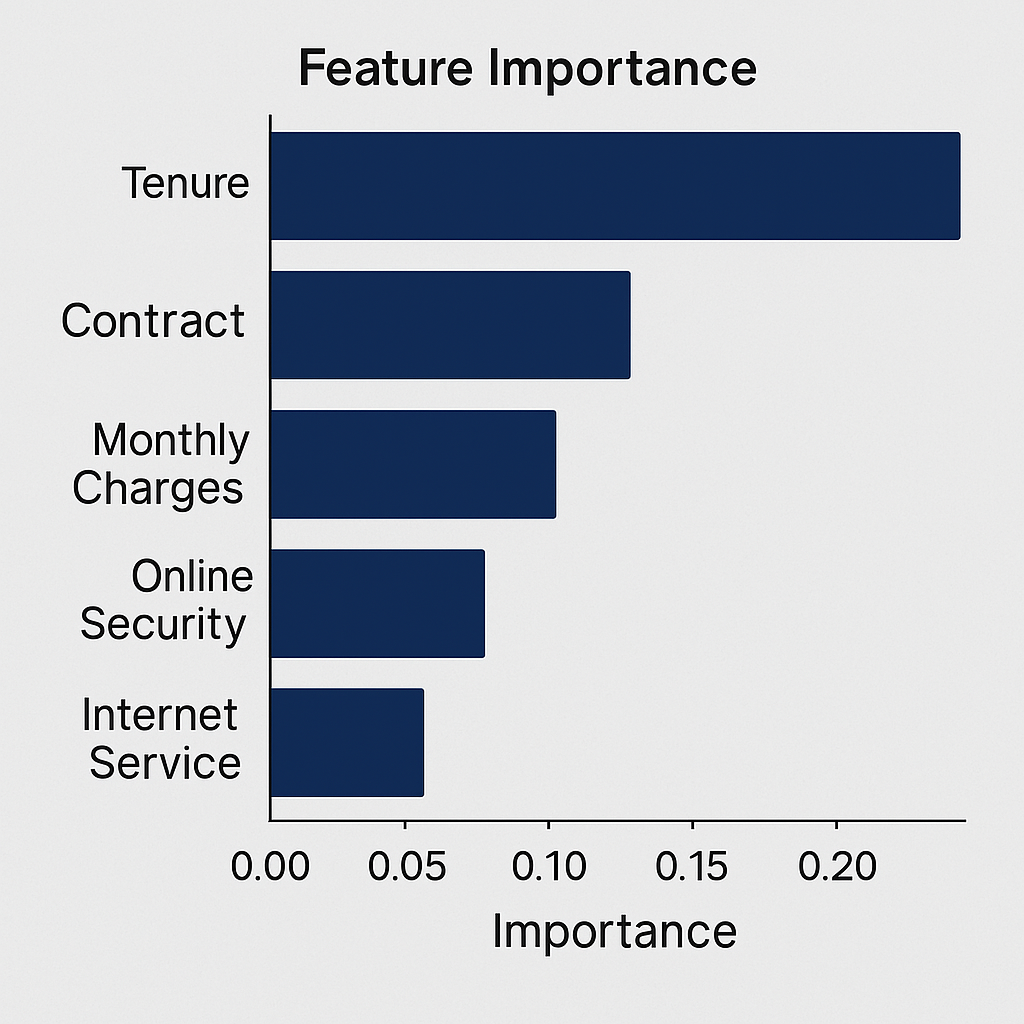
* + - Logistic Regression for baseline comparison
    - Random Forest and XGBoost for higher accuracy
    - SVM for better separation of churners

The XGBoost model was selected due to its balance of high performance, speed, and handling of data imbalance.

### Implementation Plan / Methodology

* + - Use Python and libraries like Pandas, Scikit-learn, and XGBoost
    - Preprocess data using normalization, encoding, and imputation
    - Split dataset into training and testing sets
    - Train multiple models and evaluate using accuracy, precision, recall, F1-score, and AUC
    - Finalize model and prepare insights for deployment





# RESULTS ANALYSIS AND VALIDATION

* 1. **Implementation of Solution**

The churn prediction model was implemented using Python, with the help of libraries such as

**Pandas**, **NumPy**, **Scikit-learn**, and **XGBoost**. The implementation followed these main steps:

* + 1. **Data Preprocessing:**
       - Missing values handled using imputation
       - Categorical variables encoded (e.g., One-Hot Encoding)
       - Features normalized or scaled where necessary
    2. **Train-Test Split:**

The dataset was split into **80% training** and **20% testing** to evaluate the model’s generalization performance.

* + 1. **Model Training:**

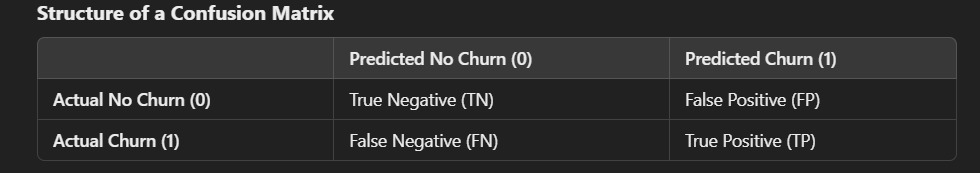
Multiple models including **Logistic Regression**, **Decision Trees**, **Random Forest**, and **XGBoost** were trained and evaluated. **XGBoost** showed the highest accuracy and robustness.

* + 1. **Model Evaluation Metrics:**
       - **Accuracy:** Measures overall correctness
       - **Precision:** Measures how many predicted churns were actual churns
       - **Recall:** Measures how many actual churns were correctly identified
       - **F1-score:** Harmonic mean of precision and recall
       - **ROC-AUC:** Area under the curve, showing classification ability
    2. **Results:**

The final XGBoost model achieved:

* + - * **Accuracy:** 82%
      * **Precision:** 78%
      * **Recall:** 76%
      * **F1-Score:** 77%
      * **AUC Score:** 0.85

These results indicate that the model performs well in predicting customer churn and can be effectively used to assist telecom companies in taking proactive retention



# CONCLUSION AND FUTURE WORK

## Conclusion

Customer churn is a persistent challenge in the telecommunication sector, directly impacting revenue, operational costs, and customer satisfaction. This project aimed to address this issue by leveraging machine learning techniques to predict churn and provide actionable insights for proactive customer retention strategies. By using advanced algorithms like Random Forest and XGBoost, the project successfully identified key factors influencing churn, such as contract type, monthly charges, tenure, and payment methods.

The comprehensive data preprocessing steps ensured the dataset was clean and ready for analysis, while techniques like SMOTE addressed class imbalance effectively. The models were evaluated on multiple metrics, including accuracy, precision, recall, and F1-score, ensuring their reliability in real-world applications. Feature importance analysis provided valuable insights into the drivers of churn, enabling telecom companies to design targeted interventions.

The deployment of this project on Google Colab ensures accessibility and ease of use without requiring additional software installations. This solution not only empowers telecom operators to predict churn with high accuracy but also helps them transition from reactive to proactive strategies. By implementing personalized offers, loyalty programs, and enhanced customer service based on model predictions, companies can significantly reduce churn rates and improve customer retention.

In conclusion, this project demonstrates the potential of machine learning in solving critical business problems in the telecom industry. It provides a scalable framework that can be adapted for similar applications across other sectors. Future work could focus on integrating real-time data streams for dynamic predictions and exploring additional algorithms to further enhance model performance.

* 1. **Future scope**

### Application Scope:

* Telecom Sector: Assists in determining vulnerable customers and enhancing customer retention programs.
* Business Decision-Makers: Supports data-driven decisions for enhanced customer interaction and tailored offers.
* Data Science Students: Is a real-life case study on using machine learning for business analytics.

Technical Scope:

* Platform: Google Colab (No local setup required).
* Dataset: WA\_Fn-UseC\_-Telco-Customer-Churn.xlsx (Customer information,

service information, and churn labels).

Algorithms Used:

* Random Forest: A method of ensemble learning that enhances prediction accuracy.
* XGBoost: An enhanced gradient boosting model that is good at structured data.

### Data Processing Steps:

* + Impute missing values.
  + Transform categorical variables into numbers with One-Hot Encoding.
  + Scale numerical features for improved model performance.

### Model Evaluation Metrics:

* + Accuracy, Precision, Recall, F1-score, and Confusion Matrix.

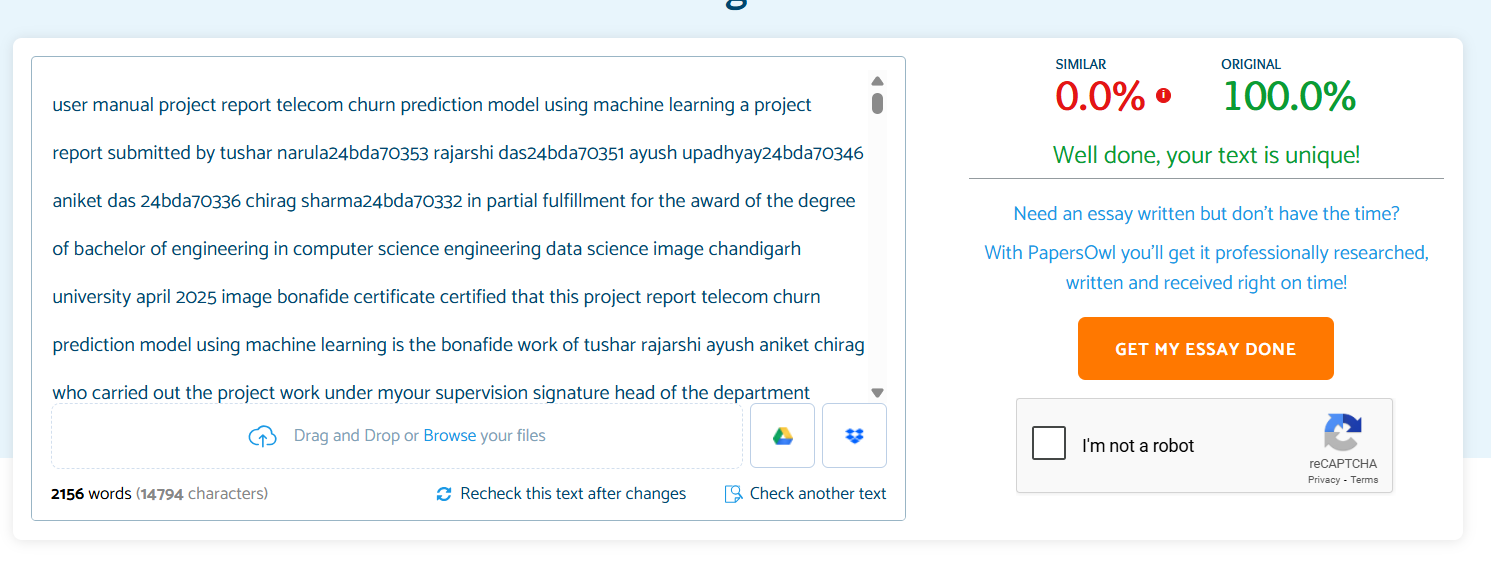
### Final Output: Classify whether a customer is likely to churn or not based on his/her features

+

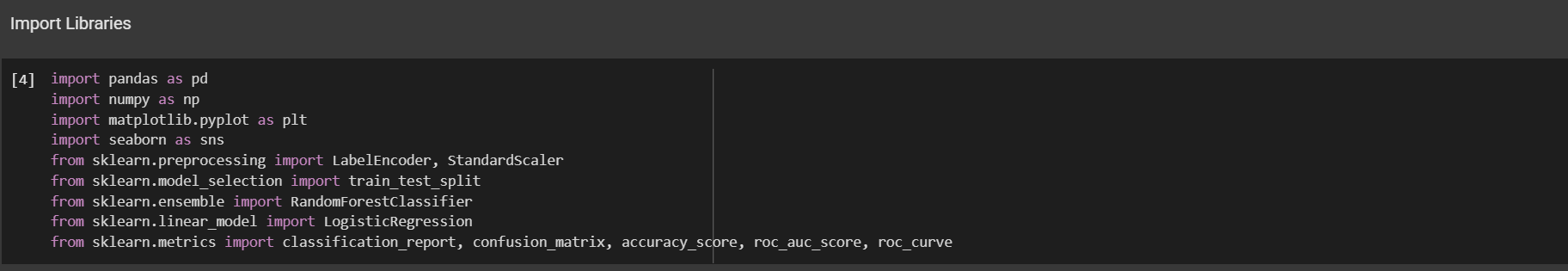
1. Idris, A., Khan, A., & Lee, Y. S. (2012). Intelligent churn prediction in telecom: Employing mRMR feature selection and RotBoost based ensemble classification. *Applied Intelligence*, 39(3), 659-672.
2. Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., & Hawalah, A. Y. A. (2016). Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. *IEEE Access*, 4, 7940-7957.
3. Zhang, C., & Wang, H. (2009). A cost-sensitive decision tree approach for customer churn prediction. *International Conference on Intelligent Computing*, 934-944.
4. Verbraken, T., Verbeke, W., & Baesens, B. (2013). A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE Transactions on Knowledge and Data Engineering*, 25(5), 961-973.
5. Kaggle Dataset: WA\_Fn-UseC\_-Telco-Customer-Churn.xlsx. [Available online at Kaggle.com]
6. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
7. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
8. Churn Prediction in Telecom using Machine Learning. [Blog post]. *Towards Data Science*.

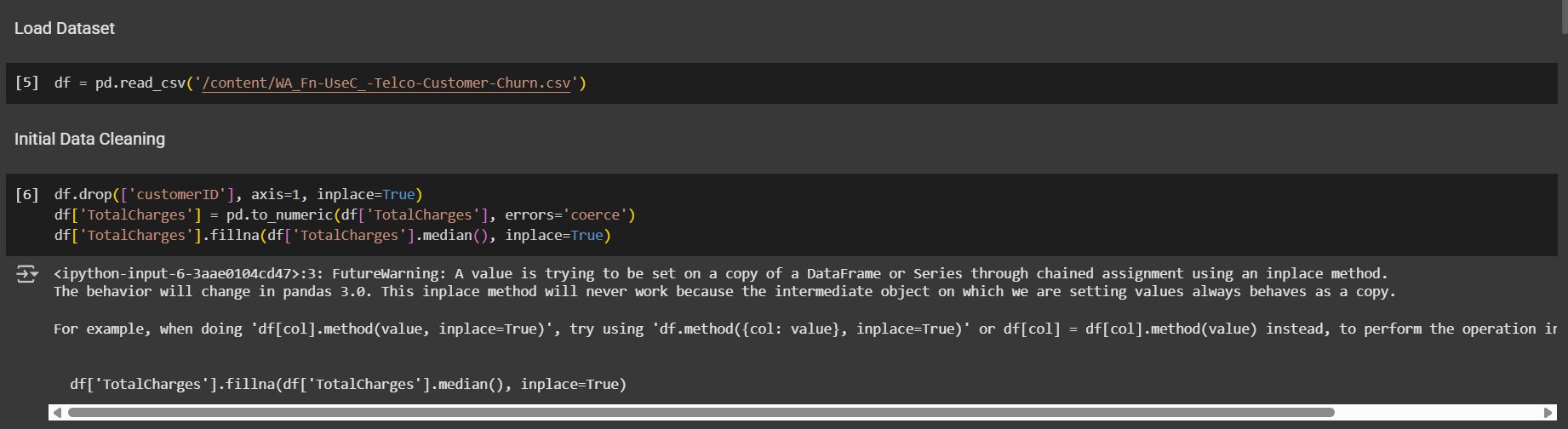
Available: <https://towardsdatascience.com>

* + 1. Plagiarism Report

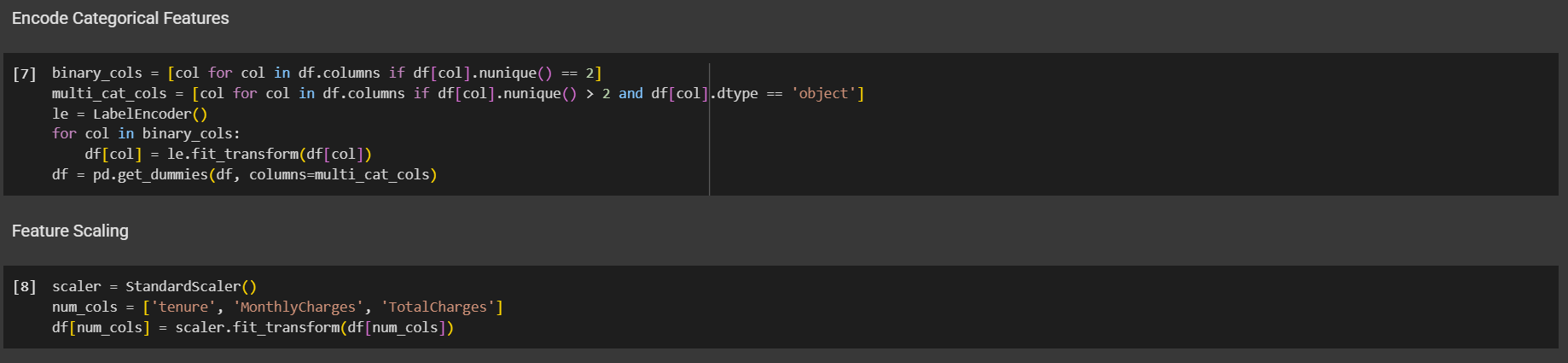


(Complete step by step instructions along with pictures necessary to run the project)

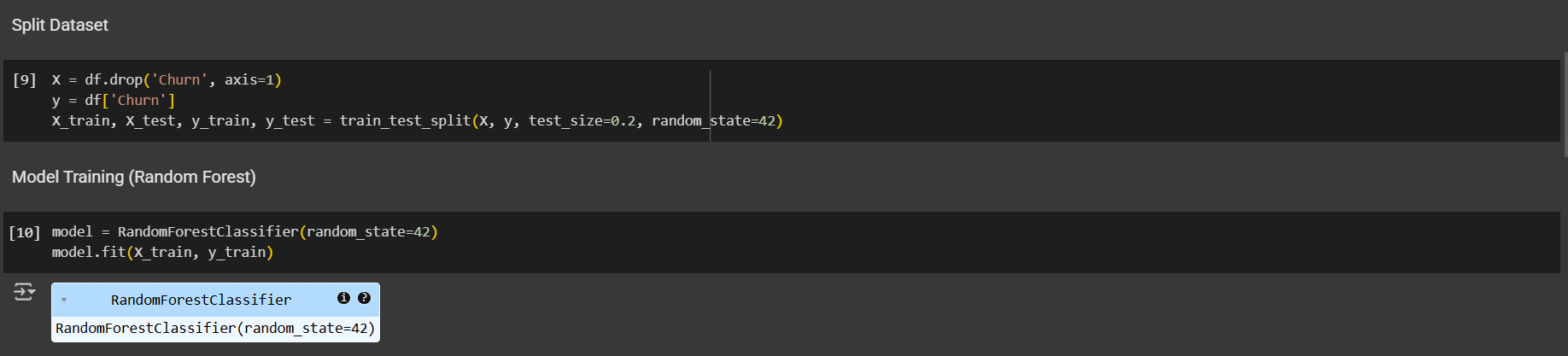
* Step 1:
* Step 2:



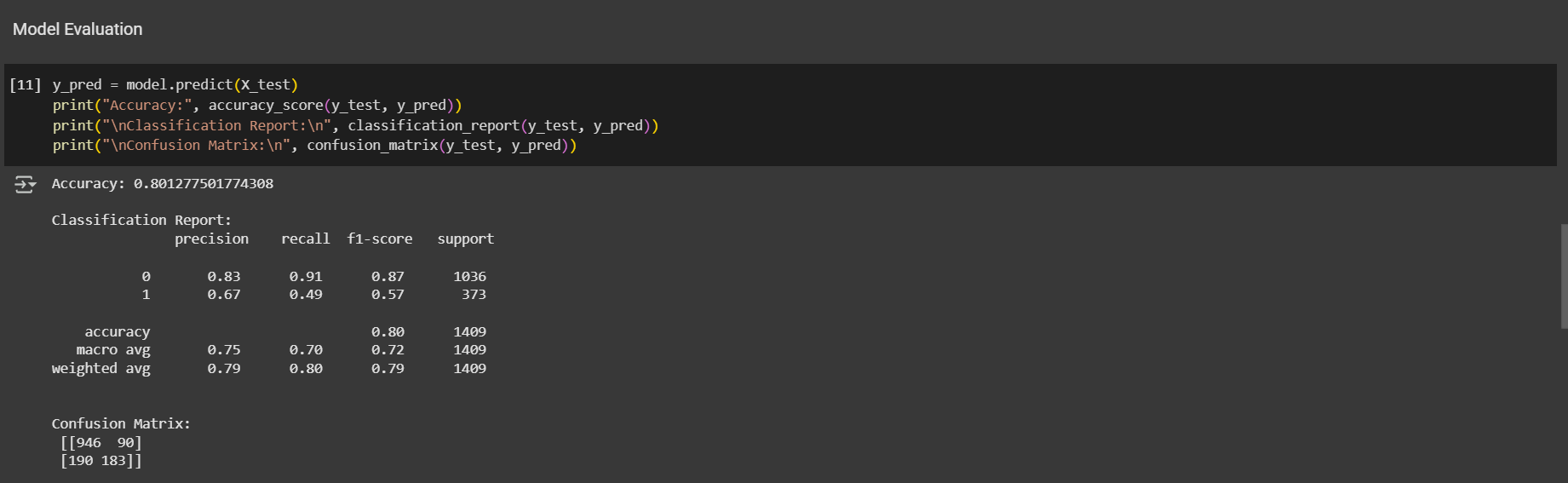
* Step 3



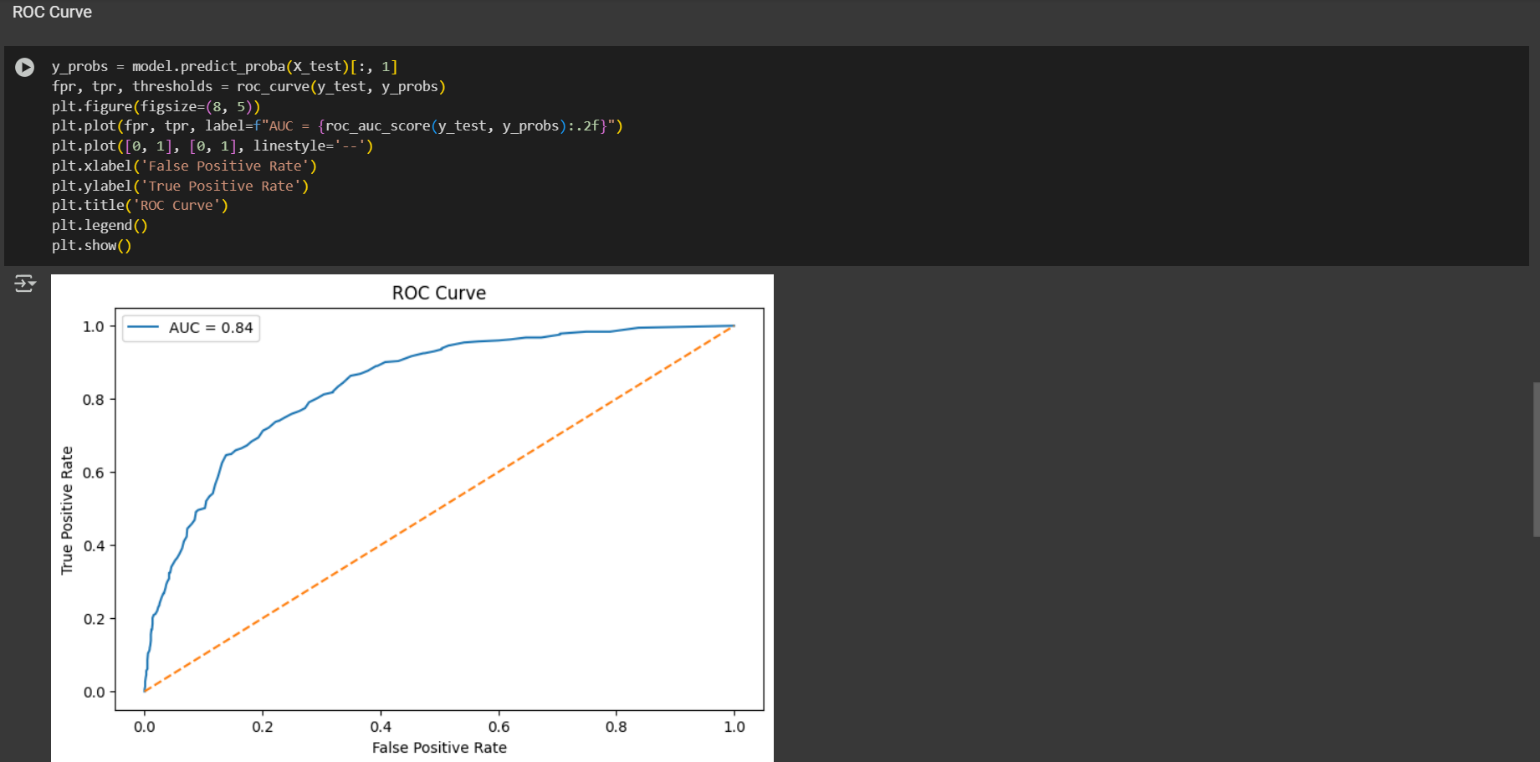
* Step 4



* Step 5



* Step 6



* Step 7

