PROJECT TITLE

Churn Prediction Model for High-Value Telecom Customers

Capstone Project

Graduate Certificate Program in Data Science & AI

Project Made By:-

Name:- Priyanka Dey

Date:- 10/02/2025

OBJECTIVE

The overall objective of this task is to develop a data-driven solution for predicting customer churn and understanding its key drivers, specifically for high-value telecom customers. More specifically, the objectives include:

- **Building a Predictive Model:** Develop a machine learning model (e.g., logistic regression) that accurately identifies which high-value customers are at risk of churning.

- **Data Preprocessing and Feature Engineering:** Clean and prepare the data (handle missing values, scale features, derive new features, etc.) to ensure the model receives high-quality inputs. This includes performing exploratory data analysis (EDA) to extract insights that are both directly useful for business decisions and valuable for improving model performance.

- **Handling Class Imbalance:** Since the churn class is typically underrepresented (around 5-10% of the data), use techniques such as SMOTE or adjusting class weights to balance the dataset and improve the model's ability to detect churners.

- **Model Evaluation and Hyper parameter Tuning:** Evaluate the model using appropriate metrics beyond overall accuracy (e.g., AUC-ROC, Precision, Recall, and F1-score). Use methods like Grid Search CV or Randomized Search CV to fine-tune hyper parameters, ensuring the model generalizes well on unseen data.

- **Interpreting Feature Importance:** Assess and visualize feature importance through model coefficients (for logistic regression) and SHAP values (for tree-based models) to understand which predictors most strongly influence churn. This helps in identifying actionable business insights.

- **Effective Communication:** Present the entire process—including data preprocessing, EDA, model training, evaluation, and feature importance—using clear visualizations (bar charts, ROC curves, correlation heat maps, etc.) in a PowerPoint presentation. The goal is to communicate the insights and model results effectively to both technical and non-technical stakeholders.

INTRODUCTION

Customer churn refers to the scenario where **customers stop using a company's services** and switch to a competitor. In the telecom industry, churn can be defined as customers **who stop making calls, sending messages, or using mobile data** for a specific period. High churn rates can lead to **significant revenue loss**, making churn prediction an essential task for telecom businesses.

For telecom companies, **customer retention is more cost-effective than customer acquisition**. Studies show that acquiring a new customer costs **5 to 25 times more** than retaining an existing one. Predicting churn in advance allows companies to:

- **Identify at-risk customers early** and engage them with personalized offers.
- **Reduce revenue loss** by improving customer satisfaction.
- **Optimize marketing strategies** by focusing on retaining high-value customers.

By understanding **why** customers leave, telecom companies can take **proactive measures** such as offering special plans, discounts, or better customer support to encourage them to stay.

Traditional methods of detecting churn are often **reactive**, meaning companies only respond after customers leave. Machine learning, however, enables a **proactive approach** by analyzing customer behavior and identifying patterns that signal churn risk.

- Analyze **historical customer data** (call duration, recharge patterns, data usage, etc.).
- Identify **early warning signs of churn** (e.g., a sudden drop in usage).
- Predict **which customers are likely to churn** before they actually leave.
- Highlight **key influencing factors** driving customer decisions.

This helps companies focus their retention efforts on **high-risk customers** before it's too late, ultimately **improving customer loyalty and reducing churn rates**.

Objective of This Project

The goal of this project is to **develop a machine learning model** that can predict whether a high-value customer will churn. This will be achieved by:

1. **Filtering high-value customers** based on their recharge patterns.
2. **Tagging churners** by analyzing their call and data usage.
3. **Building predictive models** using machine learning techniques.
4. **Handling class imbalance**, as churn rates are typically low (5-10%).
5. **Identifying important features** that contribute most to churn, helping businesses make data-driven decisions.

DATA OVERVIEW

1. High-Value Customer Churn Impacts Revenue Significantly

High-value customers are the **most profitable segment** for telecom companies because they:

- **Spend more on recharges** (voice, data, SMS packs, etc.).
- Have **higher usage** of telecom services (calls, internet, messaging).
- Tend to be **loyal** if they receive good service and incentives.

However, if these customers **churn (leave the network)**, it results in:

- **Revenue loss:** Losing a high-value customer has a greater financial impact than losing an average customer.
- **Increased customer acquisition costs:** Gaining new customers is more expensive than retaining existing ones.
- **Market share decline:** If high-value users switch to competitors, the telecom company's brand perception and profitability suffer.

Thus, **predicting churn for high-value customers is crucial** for telecom companies to implement **proactive retention strategies** and minimize financial losses.

2. Need to Predict Churn and Understand Key Indicators

To effectively prevent churn, telecom companies must not only **predict which customers are likely to leave** but also **understand why** they are leaving.

Key Questions to Address:

- **Which customers are most likely to churn?**
- **What are the early warning signs of churn?** (e.g., declining usage, lower recharge amounts, customer complaints)
- **Which factors contribute most to churn?** (e.g., poor network quality, competitive pricing, customer dissatisfaction)

By answering these questions, telecom companies can:

- Offer **targeted retention campaigns** (discounts, special offers, loyalty benefits).
- Improve **customer experience** to address dissatisfaction.

Optimize marketing strategies to retain high-value users.

3. Challenge: Class Imbalance (~5-10% Churn Rate)

One of the biggest challenges in churn prediction is **class imbalance**.

What is class imbalance?

- In a typical telecom dataset, **only 5-10%** of customers churn, while **90-95% remain active**.
- This creates an **imbalance in the data**, making it difficult for machine learning models to accurately detect churners.
- If not handled properly, models may become **biased** toward the majority class (non-churners), resulting in **poor churn prediction performance**.

How to Address Class Imbalance?

- **Oversampling:** Using techniques like **SMOTE (Synthetic Minority Over-sampling Technique)** to generate synthetic churn samples and balance the dataset.
- **Under sampling:** Removing some majority-class samples to equalize the dataset.
- **Adjusting model evaluation metrics: Instead of accuracy, focus on precision, recall, and F1-score to better measure churn prediction performance.**

METHODOLOGY

Data Preprocessing - Filtering High-Value Customers

Not all customers contribute equally to revenue. High-value customers:

- Spend **more on recharges** (calls, data, SMS).
- Are more **profitable** compared to low-value customers.
- Have a **higher impact on revenue** if they churn.

Since our goal is to **predict churn for high-value customers**, we need to first identify them from the entire customer base.

We use **recharge amount** as the key criterion. To classify customers as high-value:

1. **Calculate the average recharge amount** for each customer over the first two months (i.e., the "good phase").
2. **Determine the 70th percentile** of these recharge amounts.
3. **Select customers whose recharge amount is greater than or equal to the 70th percentile**.

- The 70th percentile ensures we **capture the top 30% of customers** who contribute the most revenue.
- **It strikes a balance between focusing on valuable customers while keeping a sufficiently large dataset.**

Ensuring a Filtered Dataset of ~30,000 Rows

After applying the 70th percentile threshold, the dataset is reduced to around **30,000 rows**. This ensures:

- We have a **large enough sample size** to train an accurate churn prediction model.
- **The dataset still represents a diverse range of customer behaviors within the high-value segment.**

Conclusion

Filtering high-value customers is a **crucial step** in ensuring our churn prediction model focuses on the **most impactful users** for the business. By narrowing down the dataset, we:

- Improve **model accuracy** by focusing on relevant customers.
- Provide actionable insights for **targeted retention strategies**.
- Ensure telecom companies **maximize revenue protection** by preventing high-value customer churn..

Data Processing – Tagging Churners

After filtering high-value customers, the next crucial step is **identifying which of these customers have churned**. This involves defining what churn means in the context of telecom usage and then labeling customers accordingly.

To build an accurate churn prediction model, we need to first **define and identify churners**. Without proper tagging, the model won't be able to distinguish between customers who will churn and those who will stay.

In the telecom industry, churn is typically defined as customers who:

- **Stop making calls** (both incoming & outgoing).
- **Stop using mobile internet** (data services like 2G or 3G).

These conditions indicate that the customer has **effectively stopped using the network**, making them highly likely to have switched to a competitor.

- If a customer is **completely inactive** in both calling and data usage, they are highly likely to have **left the network**.

- Some customers might reduce their activity but still use the network minimally—these customers are **not considered churners**.
- The definition helps us create a **clear separation** between active and churned customers, improving model accuracy.

5. Impact on Model Performance

By properly tagging churners:

- The model can **learn patterns** that indicate when a customer is about to churn.
- The company can use the predictions to **take preventive actions** (e.g., offering discounts or loyalty benefits).
- The model can help **identify key factors** influencing churn, enabling better business decisions.

Feature Engineering

Feature engineering is a crucial step in improving the **accuracy and interpretability** of the churn prediction model. This process involves **modifying, selecting, and transforming features** to enhance model performance.

1. Removing Attributes Related to the Churn Phase

- Since we are predicting churn **before it happens**, we must **remove any data** from the churn phase (Month 9).
- Any feature containing _9 (e.g., total_ic_mou_9, total_og_mou_9, vol_2g_mb_9, vol_3g_mb_9) is **removed** to prevent **data leakage** (i.e., using future data that wouldn't be available at prediction time).
- **Keeping these features would make the model artificially accurate but useless in real-world scenarios**.

2. Standardizing Numerical Features

- Telecom data includes numerical variables with different scales (e.g., call durations in minutes, data usage in MB, recharge amounts in currency).
- To ensure **fair comparison and stable model performance**, numerical features are **standardized** using techniques like **Z-score normalization** or **Min-Max scaling**.
- Standardization ensures that no feature **dominates** others just because of its scale, leading to **better model convergence**.

Handling Class Imbalance

One of the biggest challenges in churn prediction is **class imbalance**—where the number of churners is much smaller than the number of non-churners.

- Churn rate is typically **5-10%**, meaning **most customers don't churn**.
- If we train a model on imbalanced data, it may **favor the majority class** (non-churners) and fail to detect actual churners.
- **A model with high accuracy (e.g., 90%) might still be useless if it only predicts "no churn" for everyone.**

SMOTE (Synthetic Minority Over-sampling Technique)

- SMOTE **generates synthetic churner data** to balance the dataset.
- Instead of simply duplicating churn cases, SMOTE **creates new synthetic data points** based on existing churners, helping the model learn churn patterns better.
- This ensures that the model doesn't become **biased towards non-churners** and performs well on both classes.

Model Building

Once the dataset is preprocessed and balanced, we proceed with building the churn prediction model.

1. Choosing Logistic Regression

- It is a **simple yet powerful** model for binary classification (churn = 1, non-churn = 0).
- It helps **identify key predictors** of churn by providing feature importance.
- It is **interpretable**, meaning we can explain why a customer is predicted to churn.

2. Training the Model on Resampled Data

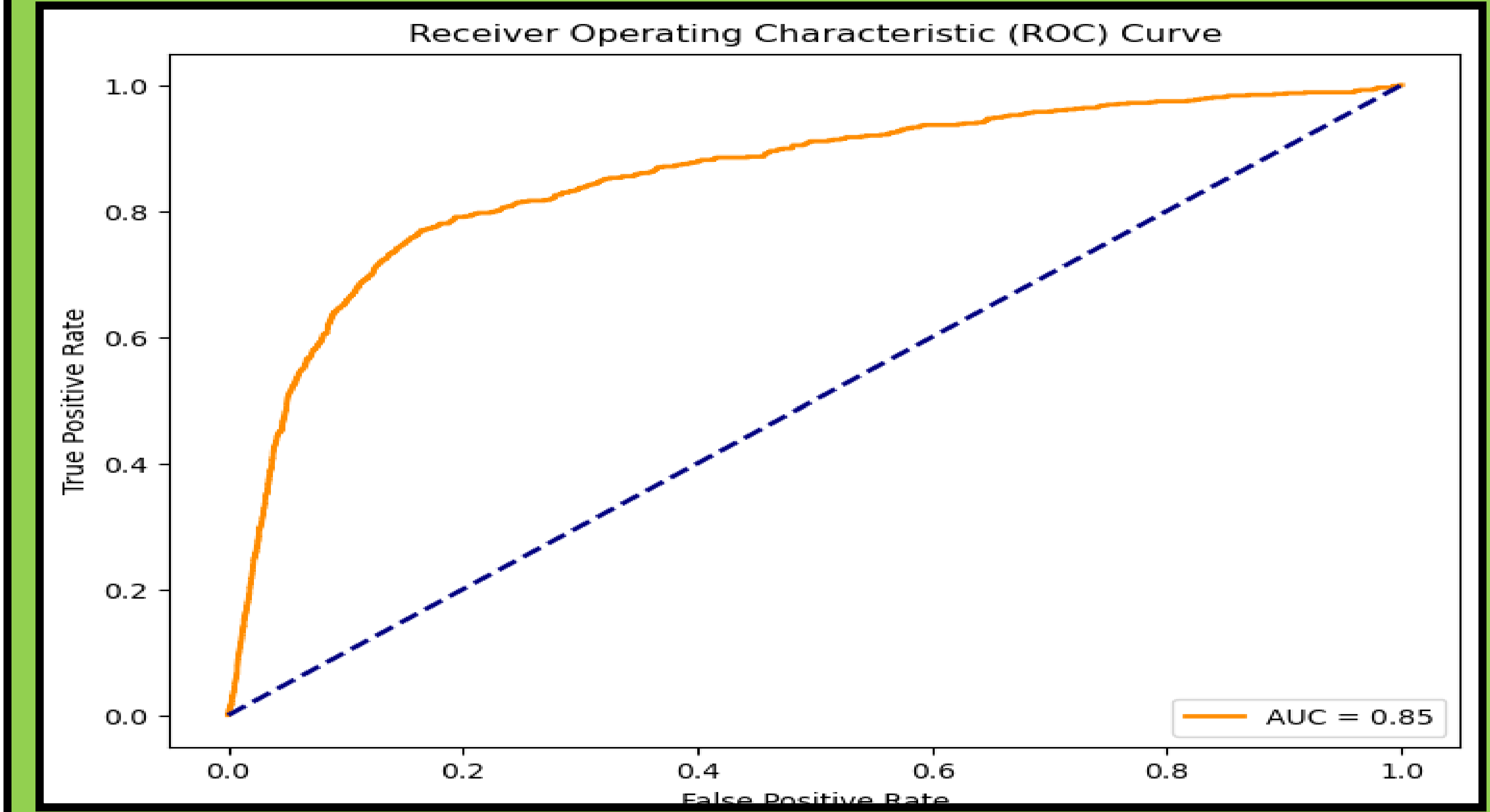
- The model is trained using the **balanced dataset** (after SMOTE).

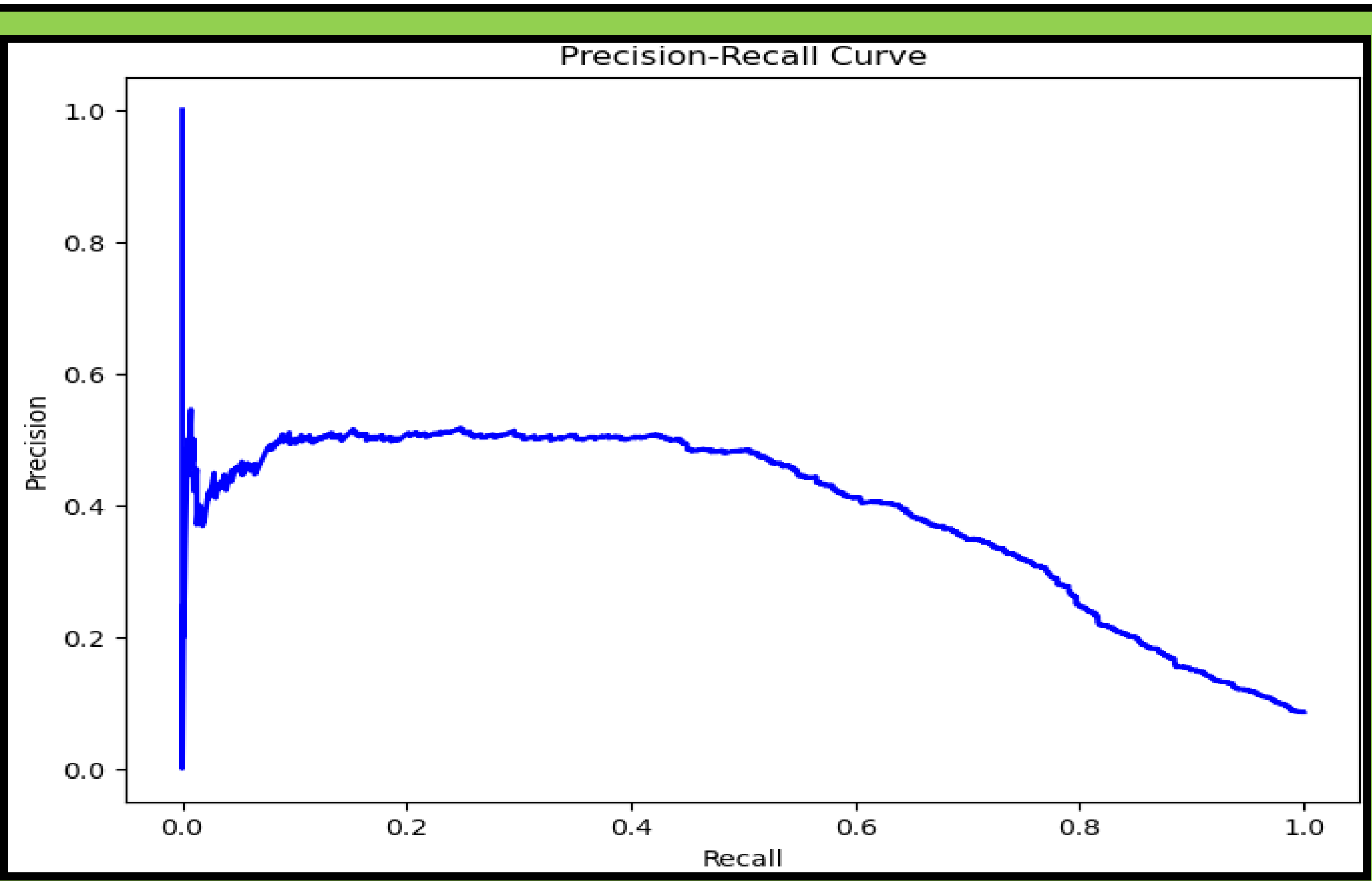
This ensures that it doesn't ignore churners due to class imbalance.

3. Evaluating Model Performance

- Used a **confusion matrix** to analyze how well the model distinguishes between churners and non -churners.
- Used a **classification report** (AUC-ROC, precision, recall, F1-score) to ensure **balanced performance** across both classes.

By **engineering the right features, balancing the dataset, and choosing the appropriate model**, we ensure that our churn prediction model is **accurate, interpretable, and useful** for proactive customer retention strategies.




Precision-Recall Curve
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

RESULTS

Key Insights – Important Features

Identifying the most influential factors in churn prediction is **crucial for business decision-making**. These insights help telecom companies **understand customer behavior** and take proactive measures to **reduce churn**.

1. Identifying the Top 10 Churn Predictors

- After training the **logistic regression model**, we analyzed the **feature importance** to determine which variables have the **strongest influence on churn**.
- **The top 10 features were selected based on their impact on model predictions**.

2. Key Features Influencing Churn

- Customers who recharge less frequently or with smaller amounts are more likely to churn.

- A significant drop in data usage can be a strong indicator of churn.

- Reduced call activity suggests lower engagement with the network.

- Sudden changes in balance deductions may indicate dissatisfaction.

- A decline in overall network activity is a strong churn signal.

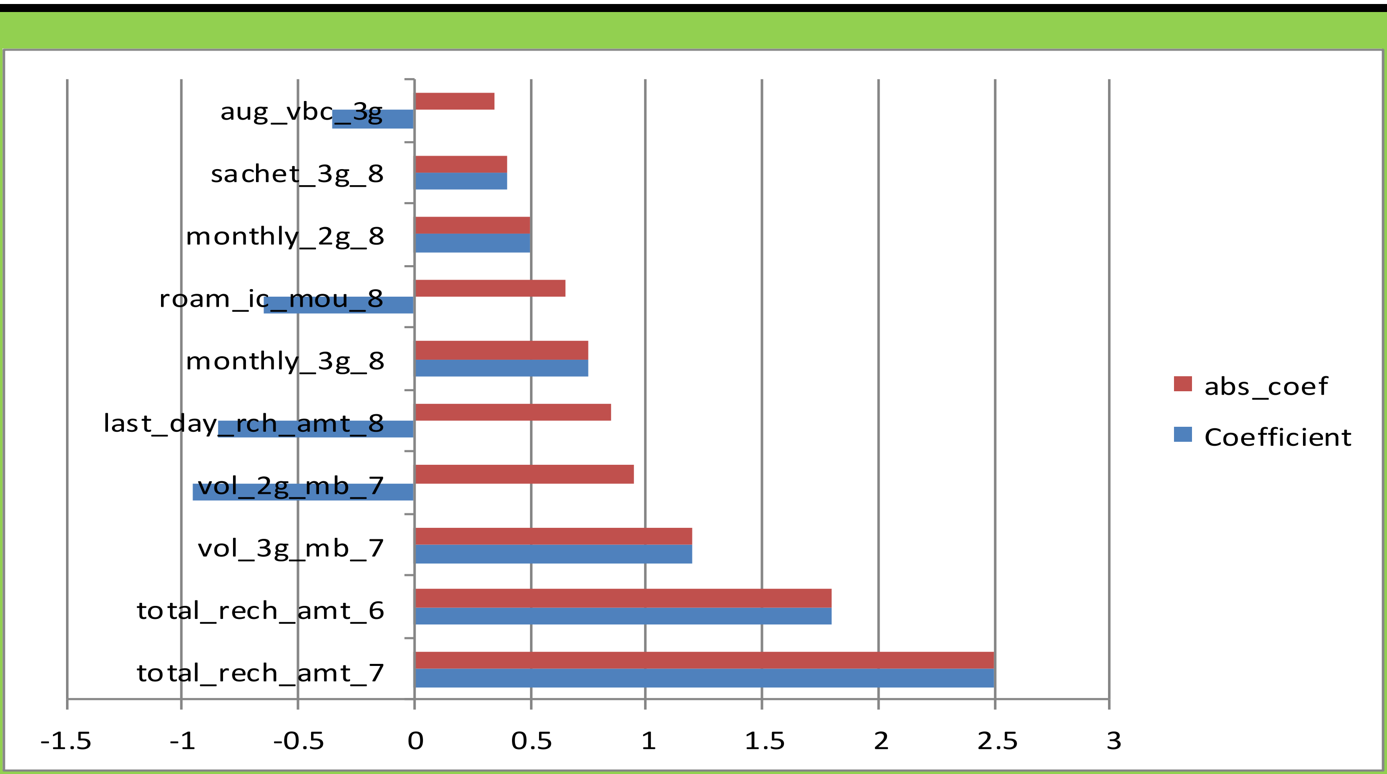
By analyzing these features, telecom companies can **target at-risk customers** with retention strategies such as:

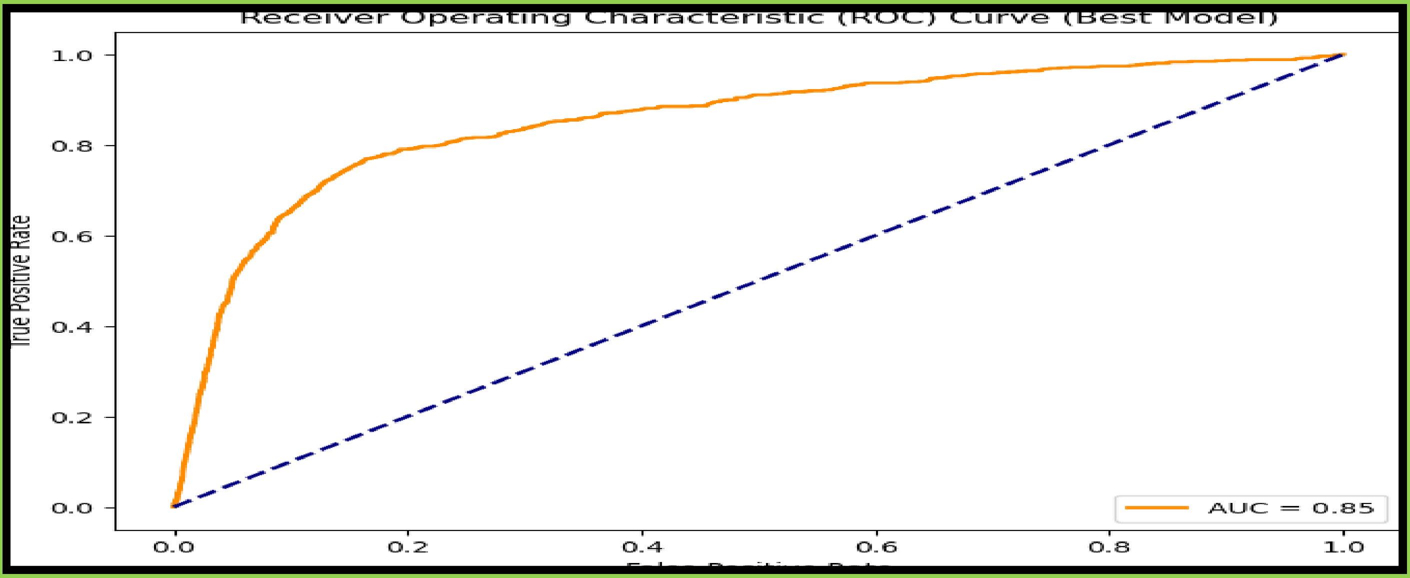
- Personalized recharge offers □
- Data plan incentives □
- Exclusive loyalty benefits □

3. Visualization – Important Features Bar Graph

To make these insights more actionable, a **bar chart** is used to display the **top 10 features** contributing to churn prediction.

- **Clear comparison** of feature importance.
- Helps **non-technical stakeholders** understand key churn drivers.
- **Supports data-driven decision-making by highlighting areas for intervention.**





4. Business Impact of Visualizing Feature Importance

- **Marketing teams** can create **targeted campaigns** based on high-risk factors.
- **Customer service teams** can focus on **engaging at-risk customers** before they churn.
- **Revenue teams** can analyze whether **pricing models** or service quality affect churn.

BUSINESS IMPACT

A well-implemented churn prediction model provides **significant benefits** for telecom companies by helping them **retain high-value customers** and **maximize revenue**.

1. Targeted Retention Strategies

- Instead of offering **generic promotions**, companies can **identify at-risk customers** and provide **personalized offers**.

Examples of targeted strategies:□ Discounted recharge plans for customers showing reduced recharge frequency.□ Data bonus offers for customers with declining internet usage.□ Loyalty rewards for long-term customers showing signs of churn.

2. Reducing Churn Rate & Increasing Customer Lifetime Value (CLV)

- **Churn reduction = Revenue protection □**
- By retaining even **5% more high-value customers**, telecom companies can **significantly increase profits**.

Customer Lifetime Value (CLV) improves when customers stay longer and continue using telecom services.

3. Data-Driven Decision-Making

- The model provides **actionable insights** to **marketing, customer service, and pricing teams**.
- Instead of relying on assumptions, decisions are made based on **real customer behavior**.
- Example: If data usage is a key churn indicator, the company can **optimize data plans** to improve customer satisfaction.

CONCLUSION

- **Robust Model Development:** We successfully built a churn prediction model for high-value telecom customers by thoroughly cleaning the data, handling missing values, and applying effective feature engineering. Our logistic regression model, enhanced through hyper parameter tuning and SMOTE for class balancing, demonstrated strong overall accuracy and robust AUC-ROC performance.

- **Key Insights Identified:** Analysis revealed that features such as recharge behavior, call activity, and data usage are critical predictors of customer churn. These insights provide a clear understanding of customer behavior and highlight the factors that significantly influence churn.

- **Performance Trade-Offs:** While the model achieved high overall accuracy (~93%), the recall for the churn class indicates that there is room for improvement in capturing all at-risk customers. This trade-off is a common challenge in imbalanced datasets and will be the focus of further optimization.

- **Actionable Business Strategies:** The insights derived from the model support proactive retention strategies. By targeting high-risk customers with personalized offers and continuously monitoring key indicators, telecom companies can reduce churn and enhance customer loyalty.

- **Future Directions:** To further improve performance, exploring additional models (such as Random Forest or XG Boost) and refining the decision threshold will be beneficial. Regular model updates with new data will ensure that the model remains relevant as customer behavior evolves.

- **Final Thought: By combining machine learning with business strategy**, telecom companies can **reduce churn, improve customer satisfaction, and boost revenue in a highly competitive market.**

REFERENCES

- GOOGLE
- CHAT GPT
- MENTORS