DATA ANALYSIS

TELECOM CHURN

Venkatraman
GDS09

INTRODUCTION

In today's competitive telecom industry, customer retention is a key challenge. This project focuses on predicting customer churn using machine learning techniques, enabling proactive intervention strategies.

We analyzed a large telecom dataset with 226 features, handling missing data, outliers, and class imbalance. Key steps included feature selection, scaling, and model optimization. Logistic Regression and Random Forest were evaluated, with Random Forest achieving 93% accuracy and a recall of 69% for churners.

The insights from this project help in identifying high-risk customers, allowing businesses to take data-driven actions to enhance customer retention.

METHODS

Data Preprocessing

- Handled missing values and outliers
- Removed irrelevant features (e.g., mobile number, circle ID)
- Converted float columns to integers for optimization

Exploratory Data Analysis (EDA)

- Identified data imbalance (low churn rate)
- Analyzed feature distributions and correlations

Feature Engineering

- Selected 15 key features using Recursive Feature Elimination (RFE)
- Addressed multicollinearity for better model stability

METHODS

Handling Class Imbalance

- Applied SMOTE (Synthetic Minority Over-sampling Technique)
- Ensured balanced training data for better model generalization

Model Training & Evaluation

- Logistic Regression (Baseline Model)
  - Accuracy: 82%, Recall for churners: 82%
- Random Forest Classifier (Final Model)
  - Accuracy: 93%, Recall for churners: 69%

Business Insights & Interpretation

- Identified key predictors of churn
- Enabled data-driven customer retention strategies

THE DATA ANALYSIS PROCESS

Data Imbalance

- Churn cases were significantly lower (5-10%), leading to biased predictions.
- Solution: Applied SMOTE and class-weight balancing to improve recall for churners.

Multicollinearity

- Some features were highly correlated, affecting model stability.
- Solution: Performed variance inflation factor (VIF) analysis to remove redundant features.

High Dimensionality

- The dataset contained 226 columns, making feature selection crucial.
- Solution: Used Recursive Feature Elimination (RFE) to retain the most relevant 15 features.

Convergence Issues in Logistic Regression

- Encountered ConvergenceWarnings due to data scaling and solver limitations.
- Solution: Scaled data using StandardScaler and increased max_iter for proper convergence.

THE DATA ANALYSIS PROCESS

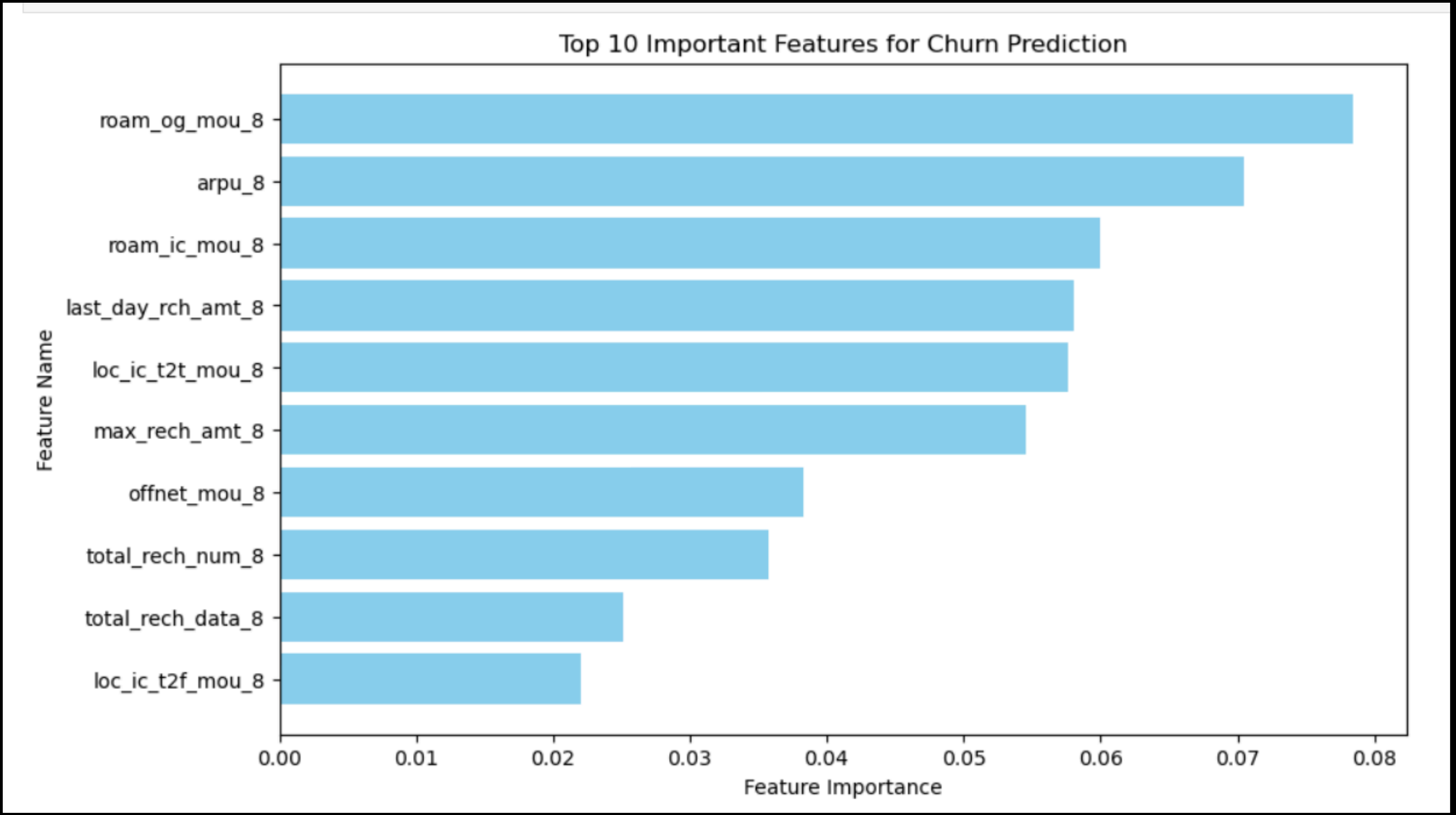
Trade-off Between Precision and Recall

- While models improved recall for churners, precision was affected.
- Solution: Chose Random Forest for a balanced approach with higher overall performance.

Business Interpretability

- Translating model insights into actionable business strategies was critical.
- Solution: Identified top churn predictors and provided visual summaries for decision-makers.

DATA VISUALIZATION



INSIGHTS AND DERIVATIONS

- The 8th month has been very important to identify customers who are likely to churn. This month is the 'action phase'.
- Identifying the usage statistics of customer after good phase is very important in the likelihood of churn.
- We need to create a model where a customers good phase is detected.
- When the usage statistics of the cuustomer is likely starting to match with the action phase, that customer is likely to churn.

The reason for churn can be :

- Offer from other network : Compare the previous recharges done with the other network to give better offers and retain customer.
- Signal problem : Compare the roaming call and see the average time spent on call. If the average is less and reason is signal, improve signal in that particular area.
- Data : Check data usage. If the person uses data or not. If data not used, use demographic factors to identify the person and give related offers. If the data usage is less or has become less over a period of time, then check for network strenght. Check if the tower has changed. Improve the signal. (Take this as a call to improve network at that area)
- Handling of the data like a stock market prediction is must. Analyse usage pattern to identify the three phases to take action during the action phase to avoid churn.

CONCLUSION

This model has given us an overview of who might likely churn, but it does not give us a name. We need to see the insights and compareit with customers.

The process has jest started. It has given an insight on who might churn but, we need to analyse the phase of the life cycle for these insights to be useful.

By creating a secong model to identify the phase of the cysle along with this model, we can accurately determine which customer is likely to churn.

THANK YOU

Venkatraman
GDS09