(CEE 590STA - Final Project Presentation)

Predicting Customer Churn in Telecommunications

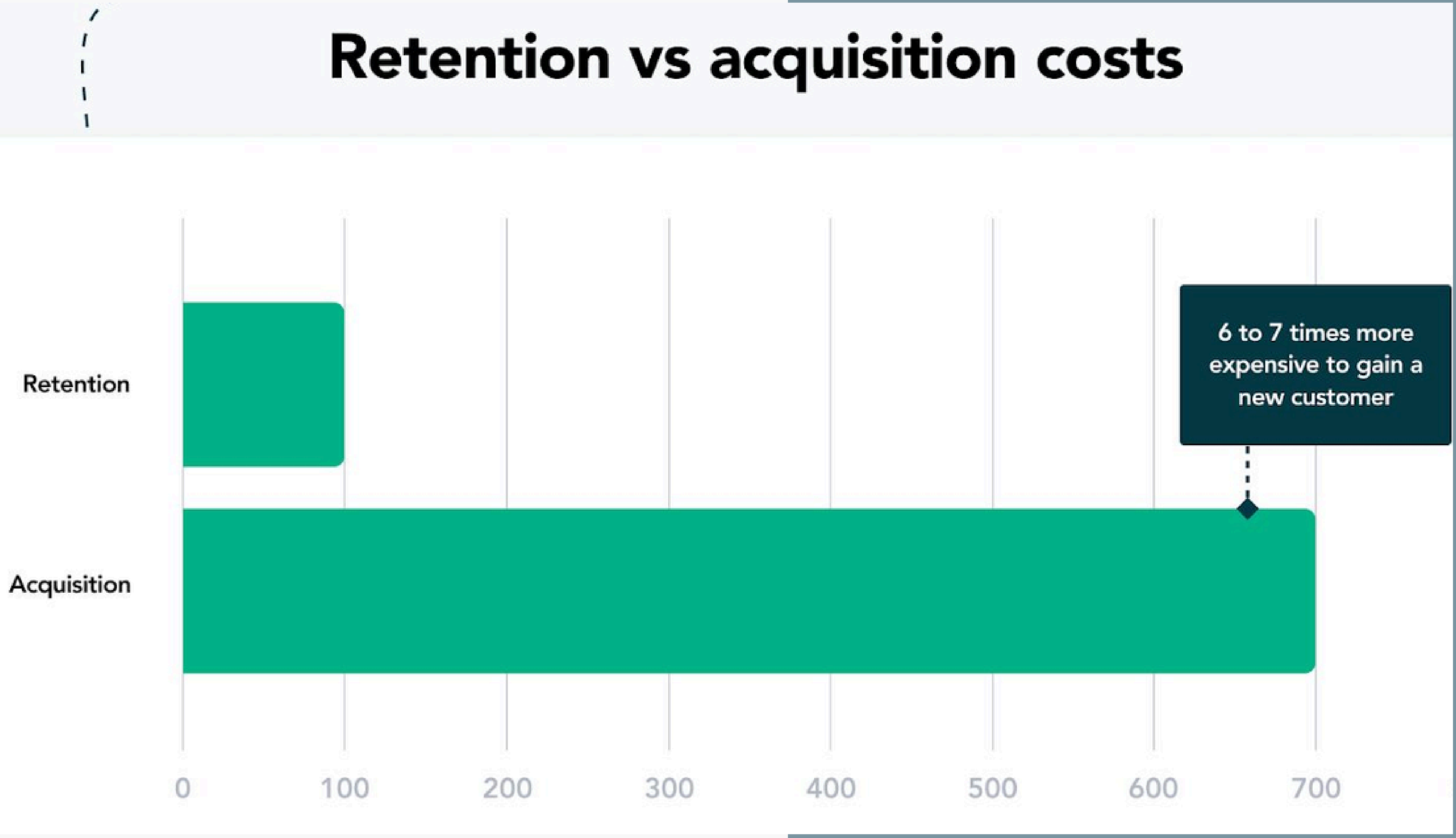
Sugun Yadla

Outline

- Project Objectives
- Data Overview & Preprocessing
- Methodology:
  - Logistic Regression
  - Gradient Boosting (XGBoost)
- Results:
  - Model Performance
  - Key Churn Drivers
- Conclusion, Limitations & Future Work

Tackling Customer Churn: A Key Business Challenge

- Customer churn is a major concern for telecommunications companies.
- Acquiring new customers is ~6x more expensive than retaining existing ones.
- Predicting churn allows for proactive retention strategies, saving costs and improving customer lifetime value.
- This project applies machine learning to understand and predict churn.

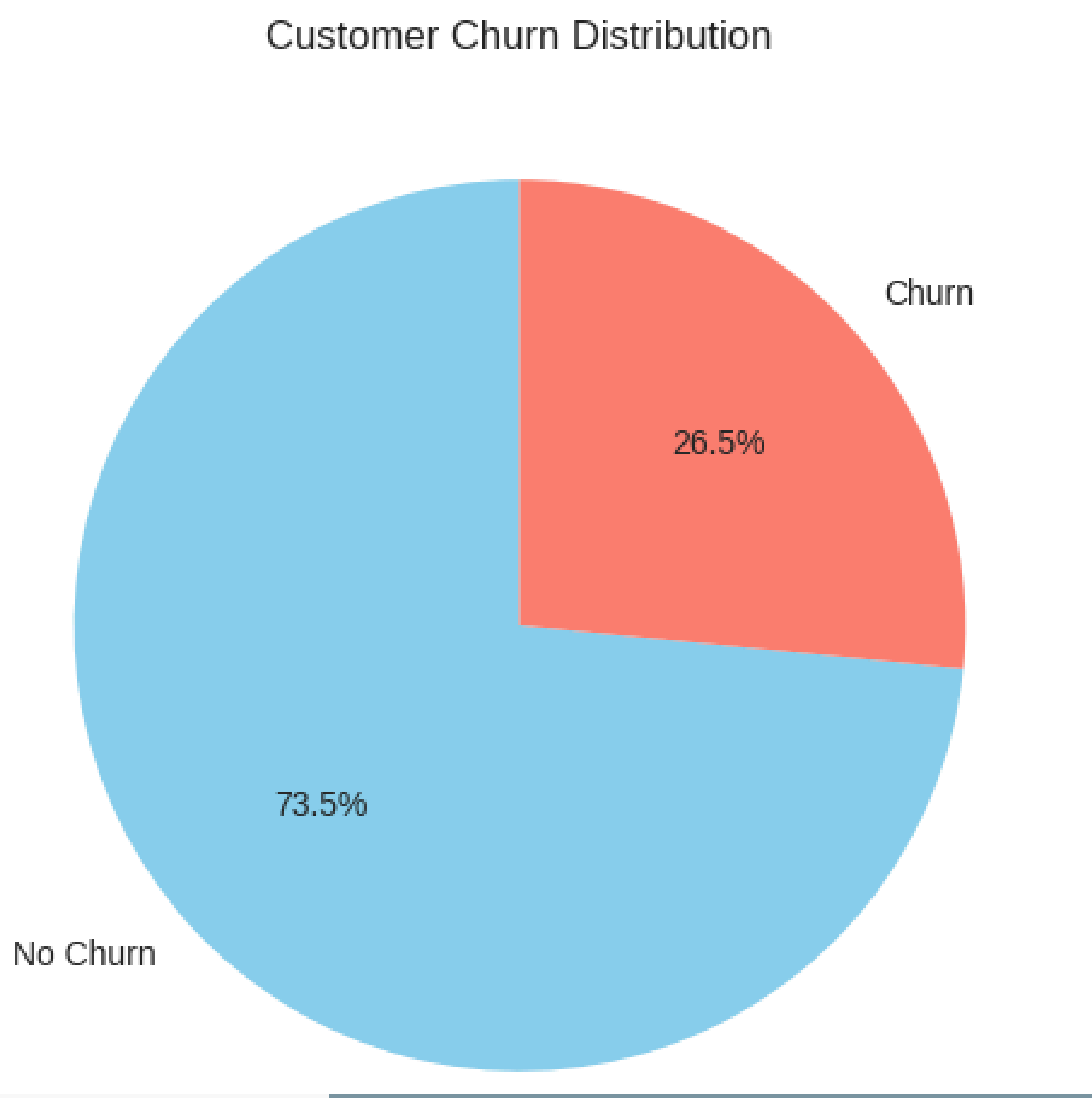


Project Objectives:

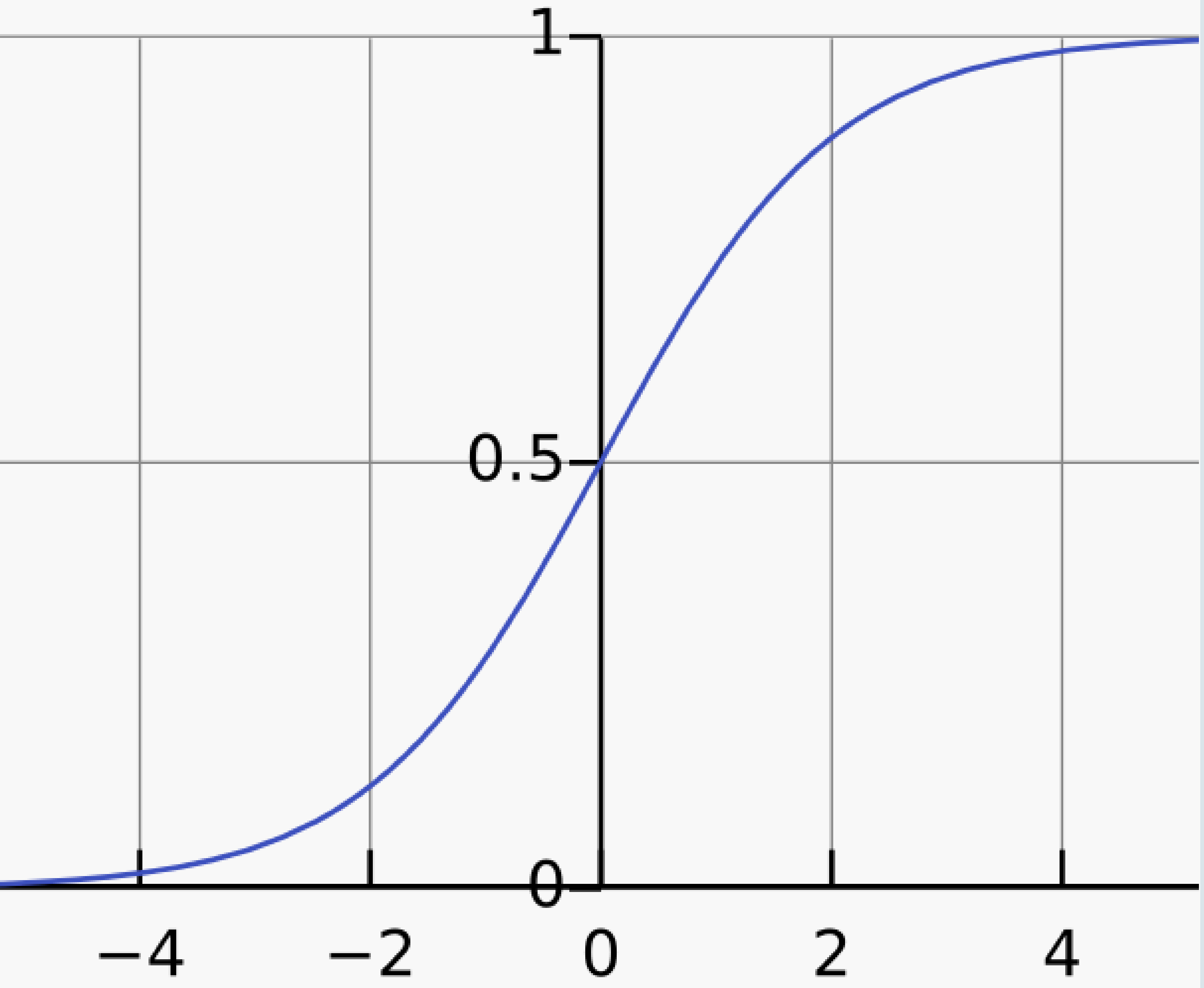
- I developed and compared machine learning models (Logistic Regression & XGBoost) for predicting customer churn.
- I evaluated the models performance using metrics like AUC, F1-Score, Precision, and Recall on unseen data. (test set)
- I then identify and analyze the key customer attributes and service factors that cause the most churn.
- This provide insights that could help create targeted retention campaigns.

The Telco Customer Churn Dataset

- Source: IBM Watson Analytics
- ~7,043 Customers, 20 Features
- Features include: Demographics, Services (Phone, Internet, TV, etc.), Contract type, Billing, Charges.
- Target: 'Churn' (Yes/No).
- TotalCharges variabble had missing values (imputed with median).
- I converted the categorical features to numerical using One-Hot Encoding.
- Scaled numerical features using StandardScaler.
- Split data: 80% Train / 20% Test (stratified).
- Target: 'Churn' (Yes/No). Class distribution shows moderate class imabalance

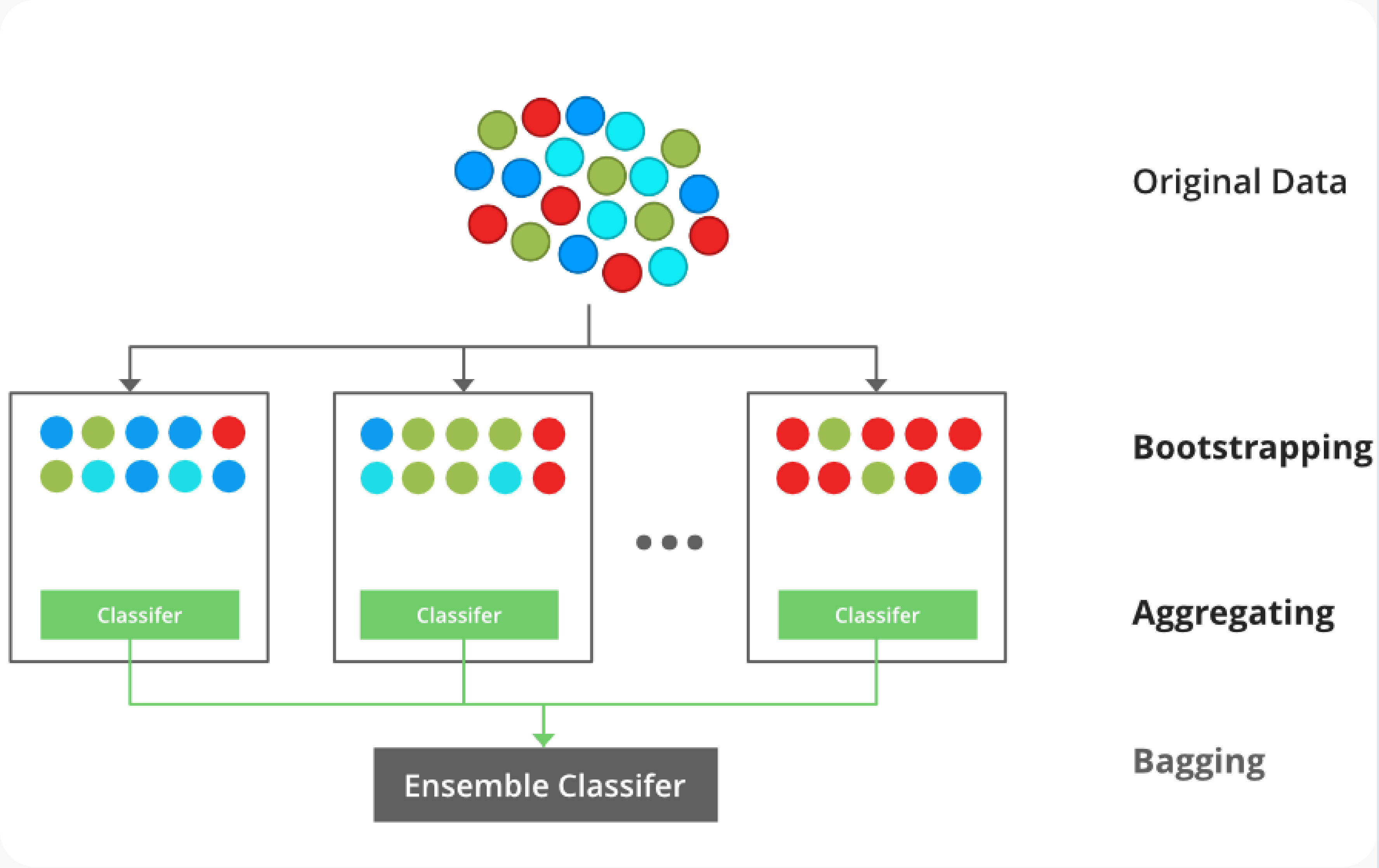


Model 1: Logistic Regression (Interpretable Baseline)



- A linear model for binary classification that predicts probability. Here is it used to predict the probability of churn.
- It uses a sigmoid function to map a linear combination of features to a probability between 0 and 1.
- Formula (Conceptual): P(Churn) = sigmoid(b0 + b1*X1 + ...)
- I chose this model first because it is fast, simple, and has highly interpretable coefficients.
- I used Scikit-learn's LogisticRegression with class_weight='balanced' to address the moderate class imbalance.
- I used One-Hot Encoding for categoricals and Feature Scaling for numericals.

Model 2: Gradient Boosting (XGBoost)

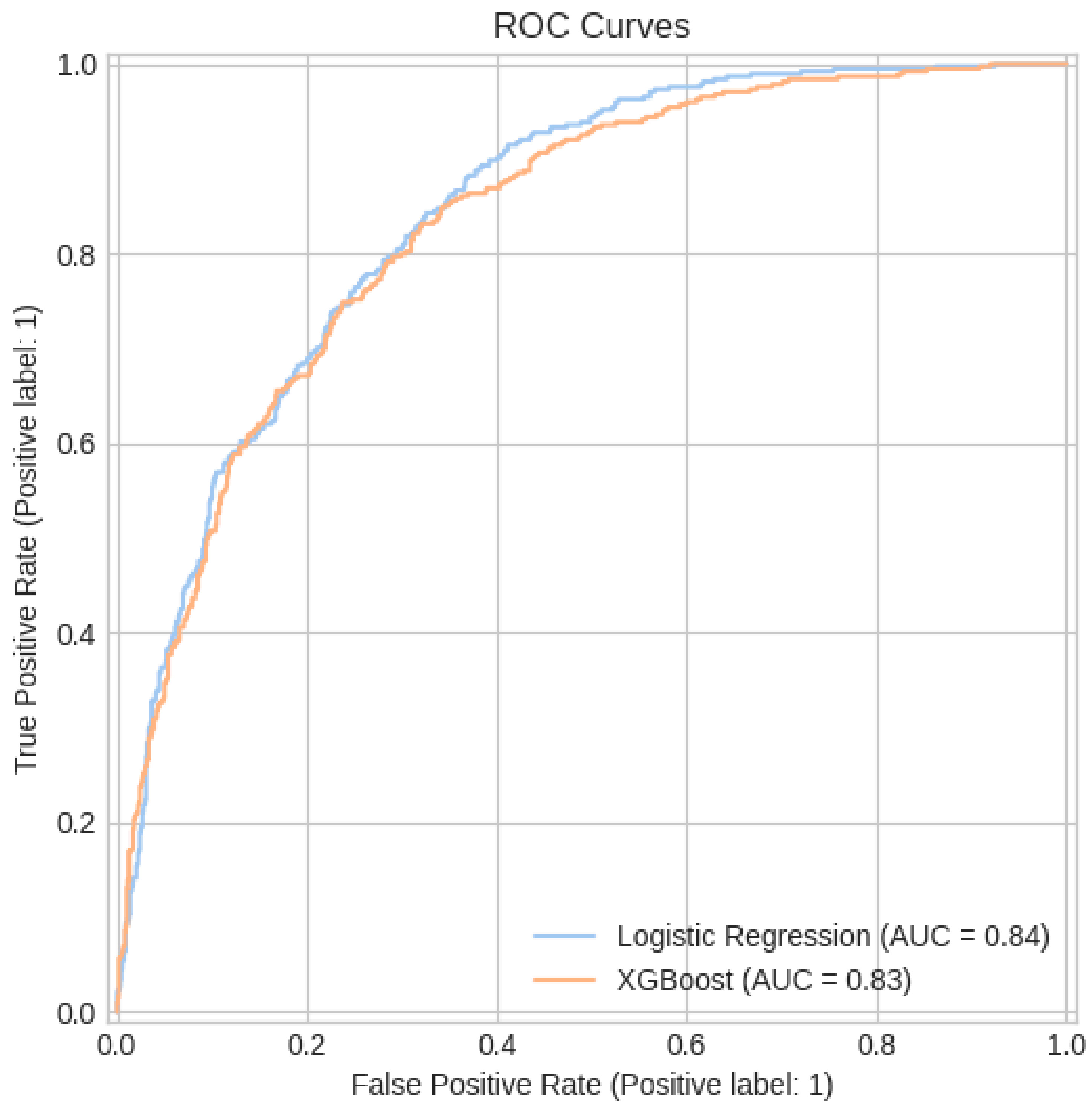


- An ensemble method that builds multiple decision trees sequentially.
- Each new tree learns to correct the errors of the previous trees, iteratively improving prediction.
- I chose this model because it has high performance potential, it can capture non-linear relationships and feature interactions and it also provides feature importance.
- I used XGBClassifier with scale_pos_weight (calculated based on training data imbalance) to manage class imbalance.
- I primarily used One-Hot Encoding for preprocessing since it is less sensitive to scaling.

Results - Model Performance

Head-to-Head: Model Performance on Test Data:

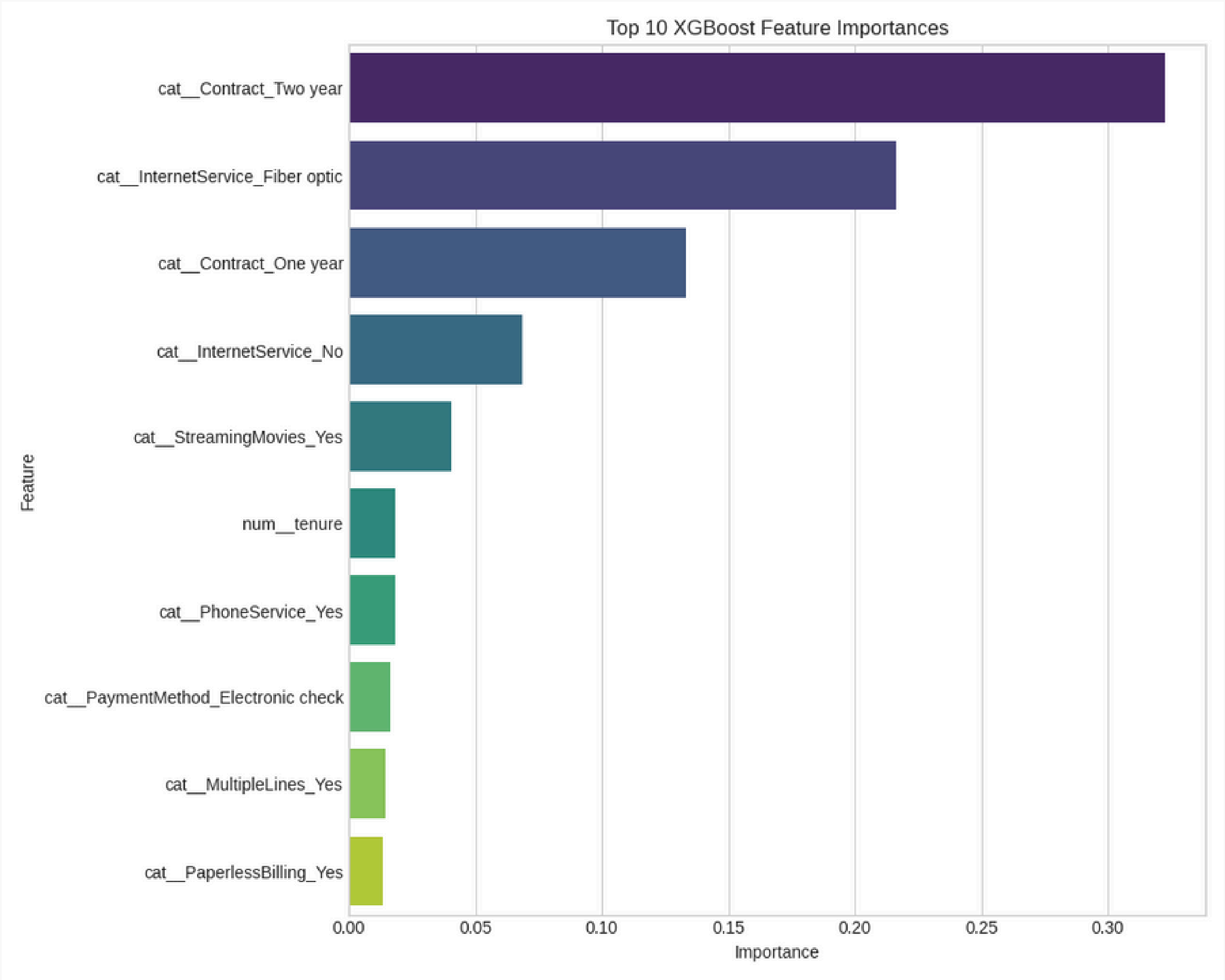
Metric	Logistic Regression	XGBoost
ROC AUC	0.8417	0.8316
F1-Score	0.6136	0.6054
Recall	0.7834	0.6952
Precision	0.5043	0.5361
Accuracy	0.7381	0.7594



Results - Key Churn Drivers

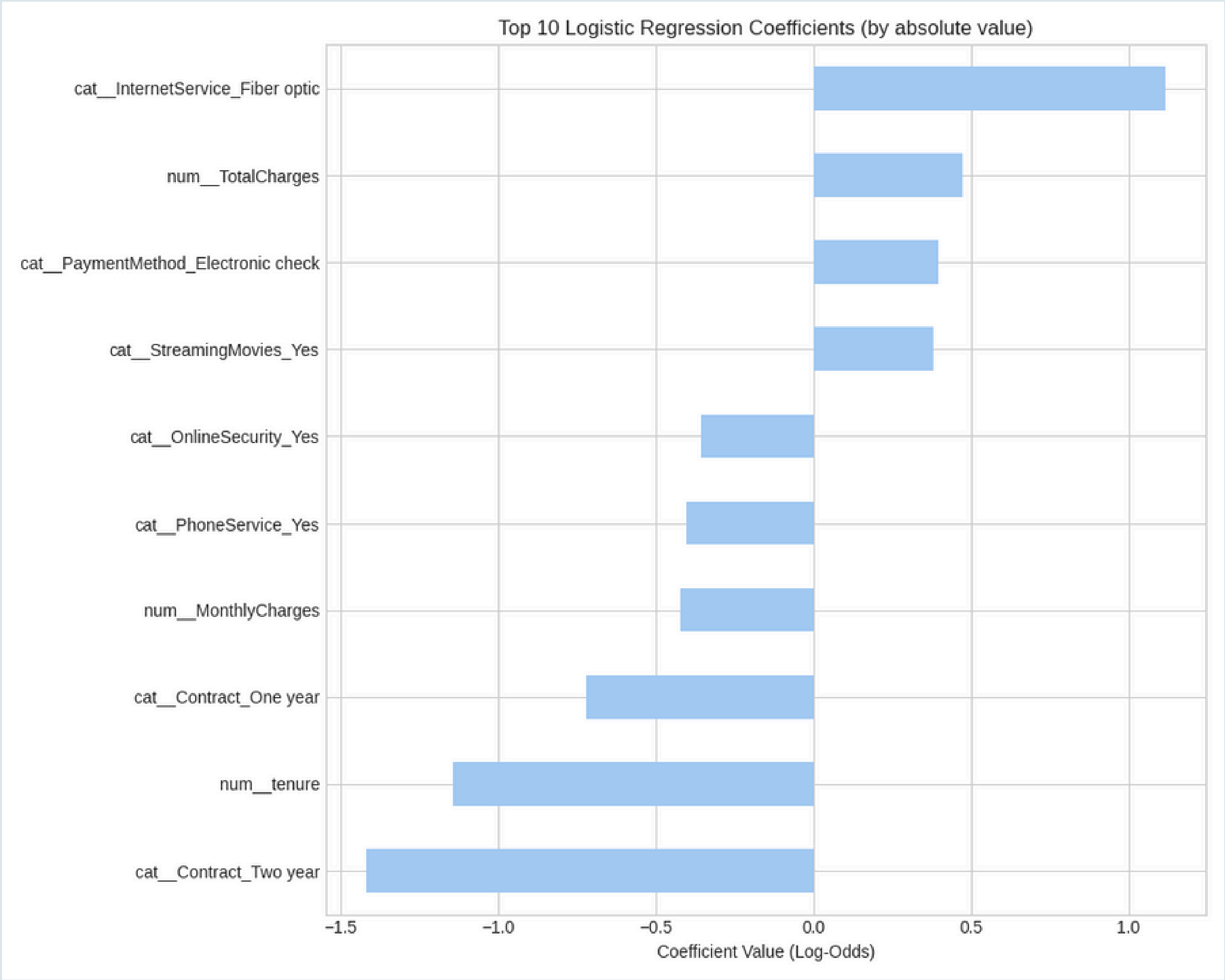
What Makes Customers Churn? Insights

XGBoost - Top Feature Importances



- XGBoost highlights Contract_Two year (likely its absence leading to churn), InternetService_Fiber optic, and Contract_One year as highly influential.

Logistic Regression - Key Coefficients:



- Logistic Regression shows factors increasing churn likelihood are InternetService_Fiber optic, which has the strongest positive impact, PaymentMethod_Electronic check and Higher TotalCharges
- The factors decreasing churn likelihood are one and two year contracts(highest negative impact) and tenure

Combined Insights:

- Both models point to contract type as crucial, month-to-month customers, are often the reference category when 'One year' and 'Two year' contracts reduce churn, are at higher risk.
- Subscribing to Fiber Optic Internet service consistently appears as a factor associated with higher churn risk across both models, potentially due to the price of the service.
- **Longer tenure and longer-term contracts are key to retention.**

Conclusion, Limitations & Future Work

Summary:

- I built and evaluated Logistic Regression and XGBoost models for Telco churn prediction, which allowed us to understand the causes behind the moderate class imbalance.
- The Logistic Regression performed very strongly, achieving an AUC of 0.842 and an F1-Score of 0.614, with a Recall of 0.783. It slightly outperformed XGBoost (AUC 0.832, F1 0.605) in these areas.
- The key churn drivers were contract type (month-to-month risk), fiber optic internet subscription, and payment method (electronic check). Longer tenure and contracts significantly reduce churn.
- It shows that more complex models aren't always better and that a well-tuned, interpretable model like Logistic Regression can be very effective, especially when imbalance is handled

Limitations:

- The dataset is static a real-world churn is dynamic, over a period of time
- External factors such as competitor actions are not modeled, even though they potentially could be significant driving factor for churn in the first place.

Future Work

- There is room for further hyperparameter tuning for both models.
- I could explore other algorithms or ensemble techniques.
- A deeper dive into feature interactions, especially around TotalCharges and service types.
- Developing and test targeted retention strategies (e.g., for Fiber Optic customers on month-to-month contracts) would potentially be the next step in reducing churn.

Thank you