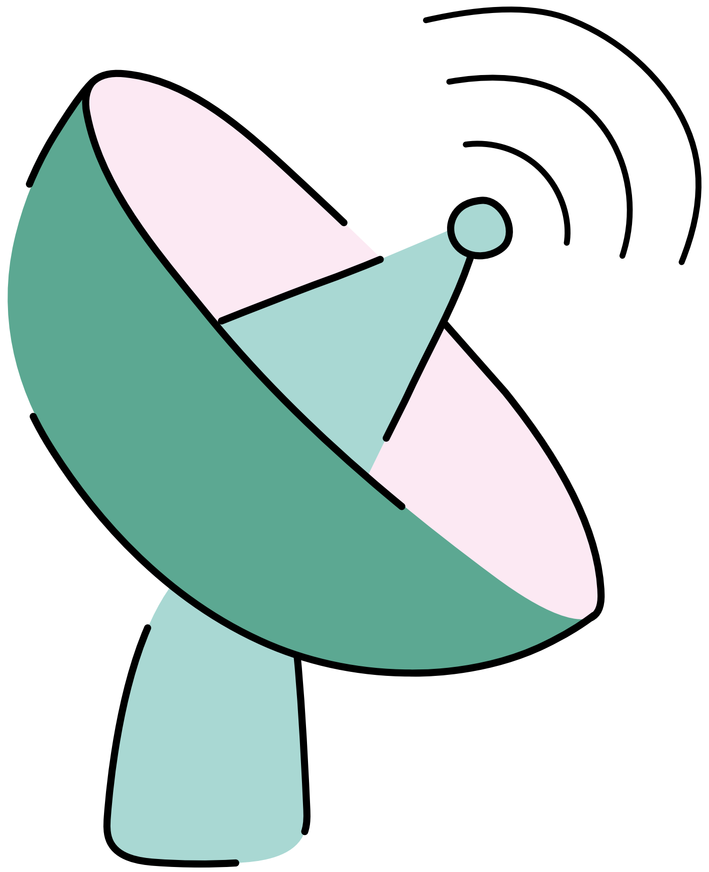
Customer Relationship Prediction



TEAM FLIPPING-A-COIN

Shambhavi Aggarwal

Lakshay Chawla

Devansh Shrestha

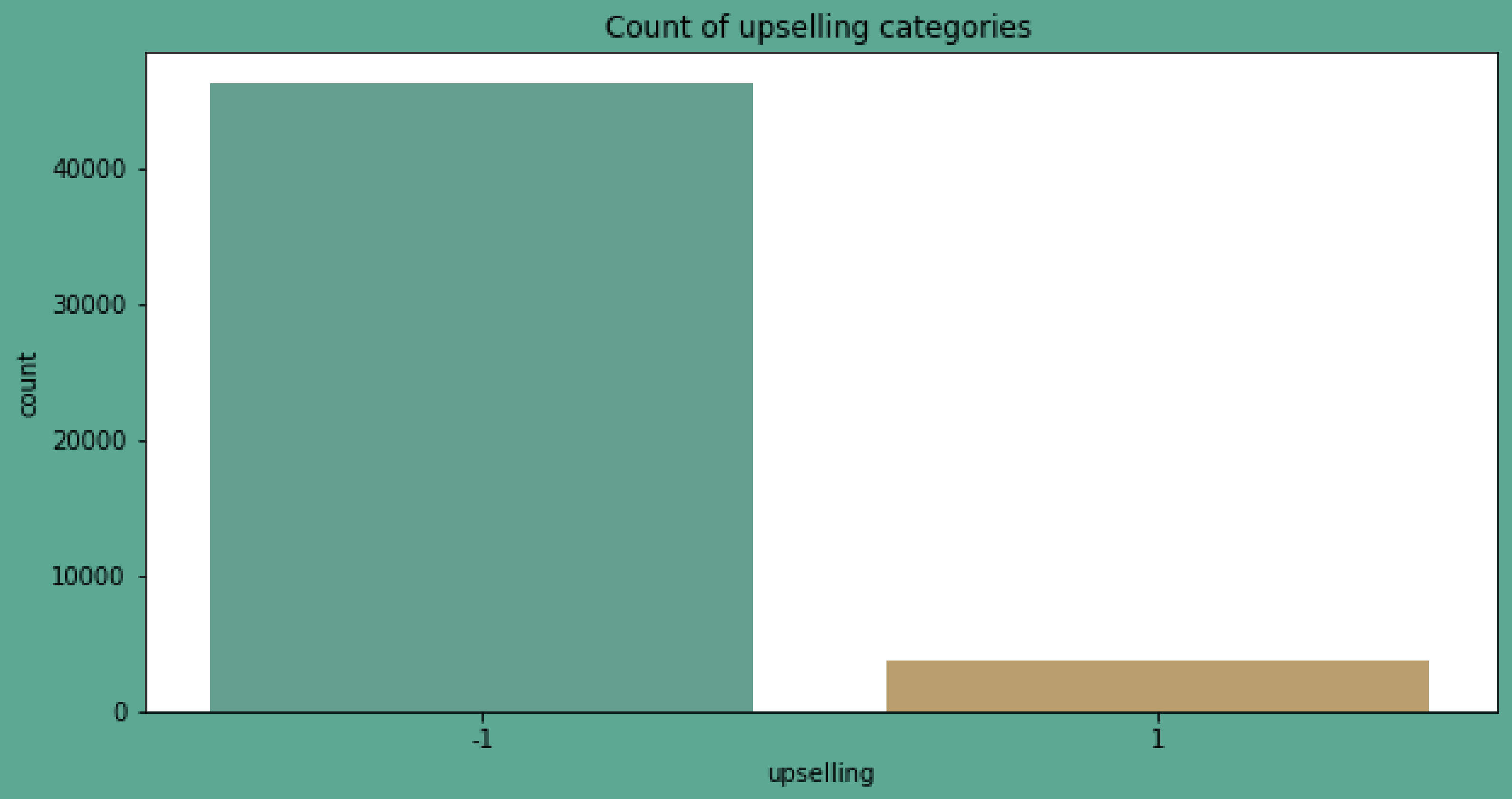
Rommel Jalasutram

The Problem.

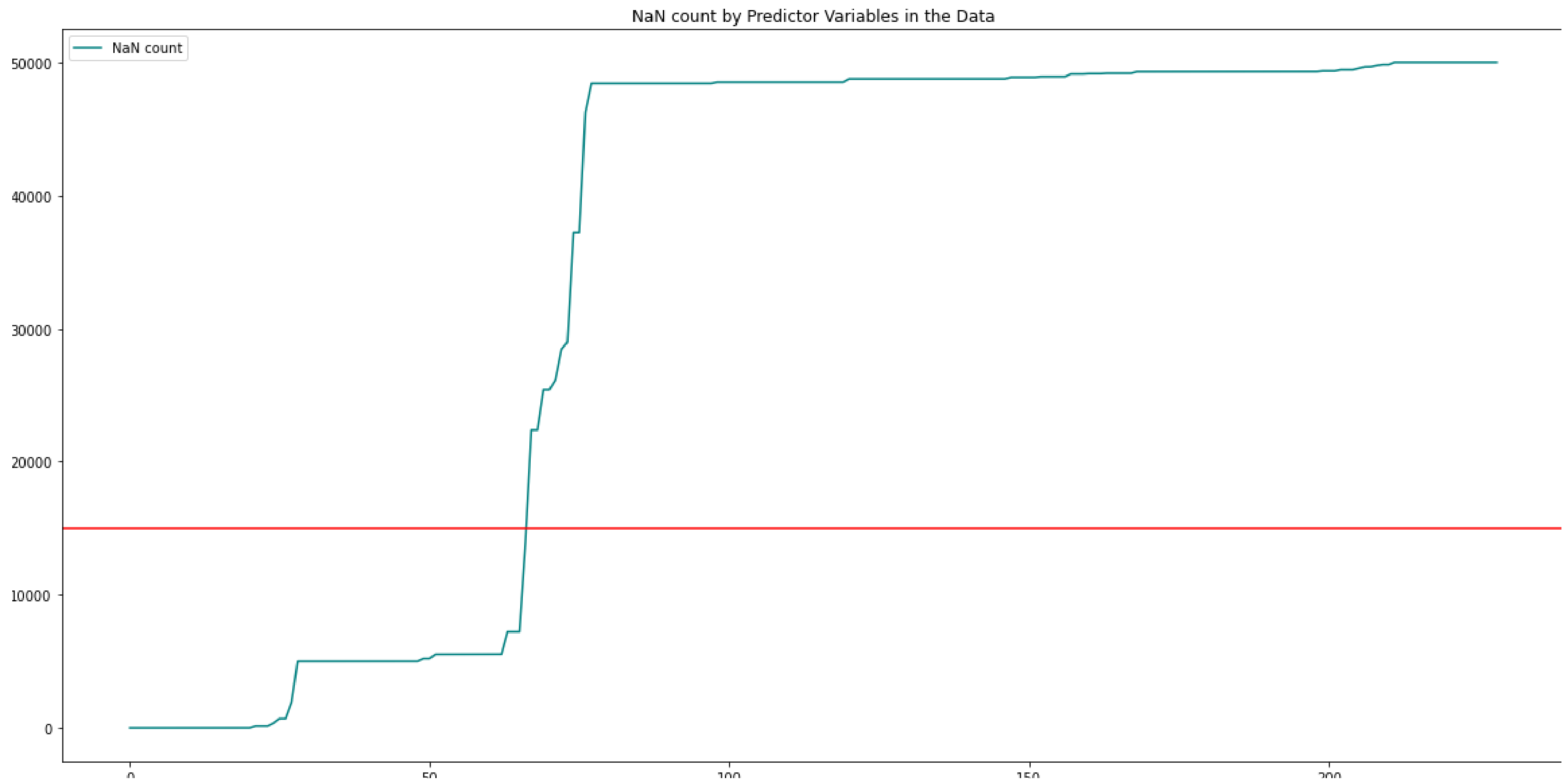
Orange, a French Telecom Company has collected data of their customer activities. Based on the data, we are predicting the propensity of customers to switch providers (churn), buy new products or services (appetency), or buy upgrades or add-ons proposed to them to make the sale more profitable (up-selling).

The Dataset.

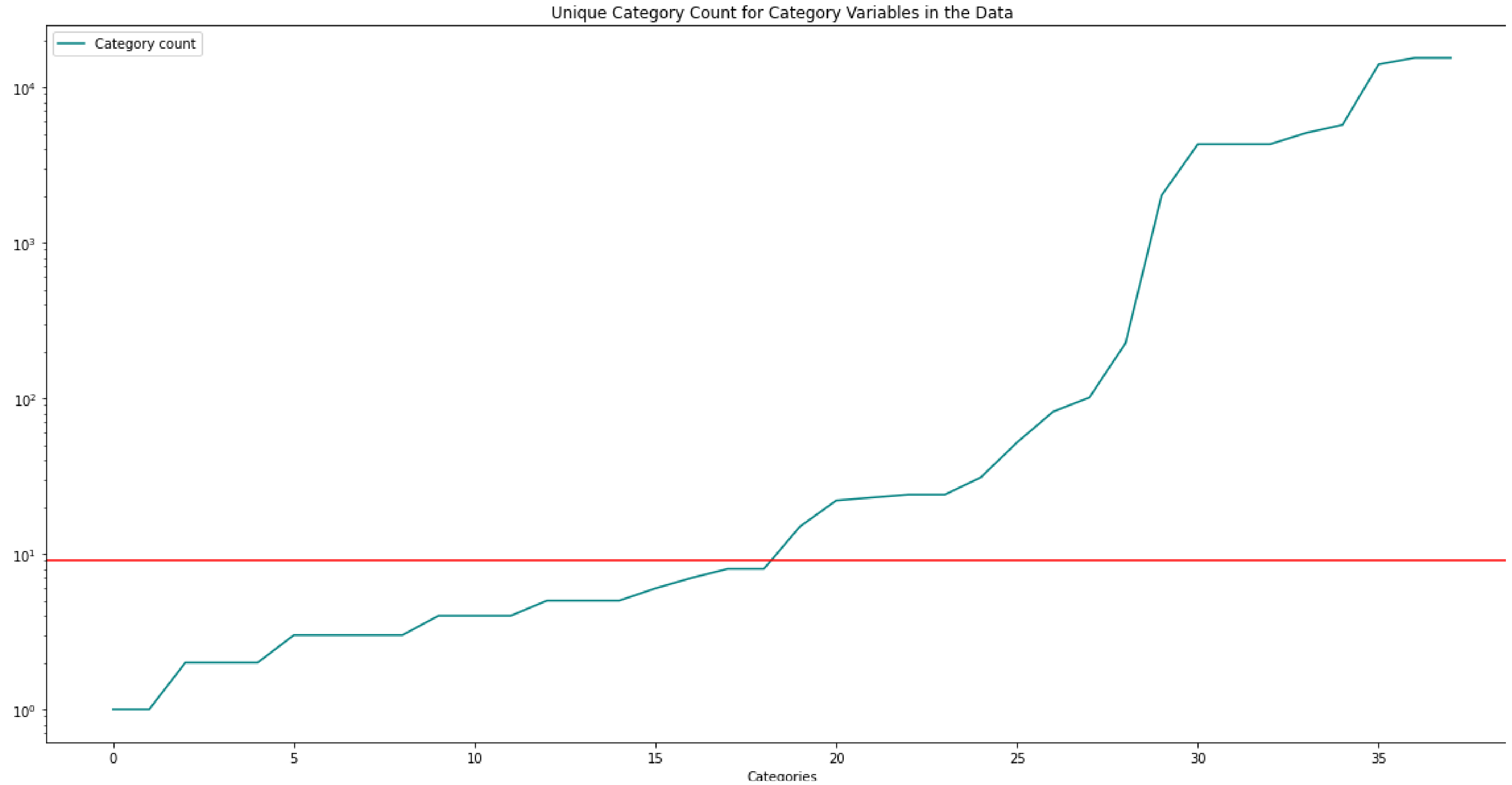
- The smaller version of the dataset contains 230 features and 50000 observations.
- 190 numerical variables.
- 40 categorical variables.
- Target variables heavily unbalanced.



Exploratory Data Analysis.



- Majority predictor columns contain > 30% NaN values.
- NaN for Numerical columns imputed with median.



- Majority category columns contain >= 9 unique categories.
- NaN for Categorical columns imputed with max category.

Methodology.

- Split data: training and testing using stratify on target variable.
- Combine useful features based on:
  - Information Gain
  - Random Forest Feature Importance.
- Use GridSearchCV to get best hyper-parameters.
- Fit the following models:
  - Logistic Regression
  - Decision Trees
  - Vanilla Random Forests
  - Balanced RF
  - RF with up-sampling
  - RF with down-sampling
  - Adaboost
- Predict and compare models.
- Analyse Permutation Importance.

Results.

- Upselling
  - Logistic Regression: Lowest F1 and AUC score.
  - AdaBoost: Highest F1 and high AUC score.
- Appetency
  - Logistic Regression: Lowest accuracy, low F1 & highest AUC score.
  - AdaBoost: Highest accuracy, zero F1, and lowest AUC score.
- Churn
  - Logistic Regression: Low accuracy, low F1, and lowest AUC score
  - AdaBoost: Highest accuracy, zero F1, and high AUC score.

Discussion.

- Upselling
    - Tree models gave best performances.
    - Adaboost performed the best.
    - Var 126 had highest permutation importance.
- Appetency
    - None of the models performed well.
    - Improve performance by balanced datasets, resampling.
    - Can be improved by Support Vector Machines, XGBoost.
- Churn
    - Only 2% target variables positive.
    - Improve performance by balanced datasets, resampling.
    - Var 126 had highest permutation importance.

Conclusion.

- Tree models performed well.
- Data highly imbalanced.
- Models highly biased towards majority class as a result.
- Grid search while useful is computationally expensive.

Future Scope.

- Train with fewer predictor variables.
- Improve performance by balanced datasets, resampling.
- Try SVM and XGBoost on these datasets.

Thank you.