Credit Payment Analysis and Default Detection

Bhargav Naik, Garrett Wadley, Grace Chiu, Harshika Shete, Luiz Roma, Yixuan Jin

Data Source

Data: Default Payments of Credit Card Clients in Taiwan April 2005 to September 2005

Source: Kaggle credit card default detection

Target label: binary, default next month as 1, 0 otherwise

Categorical Data: sex, education, marriage, age, payment statuses (on time or number of months late)

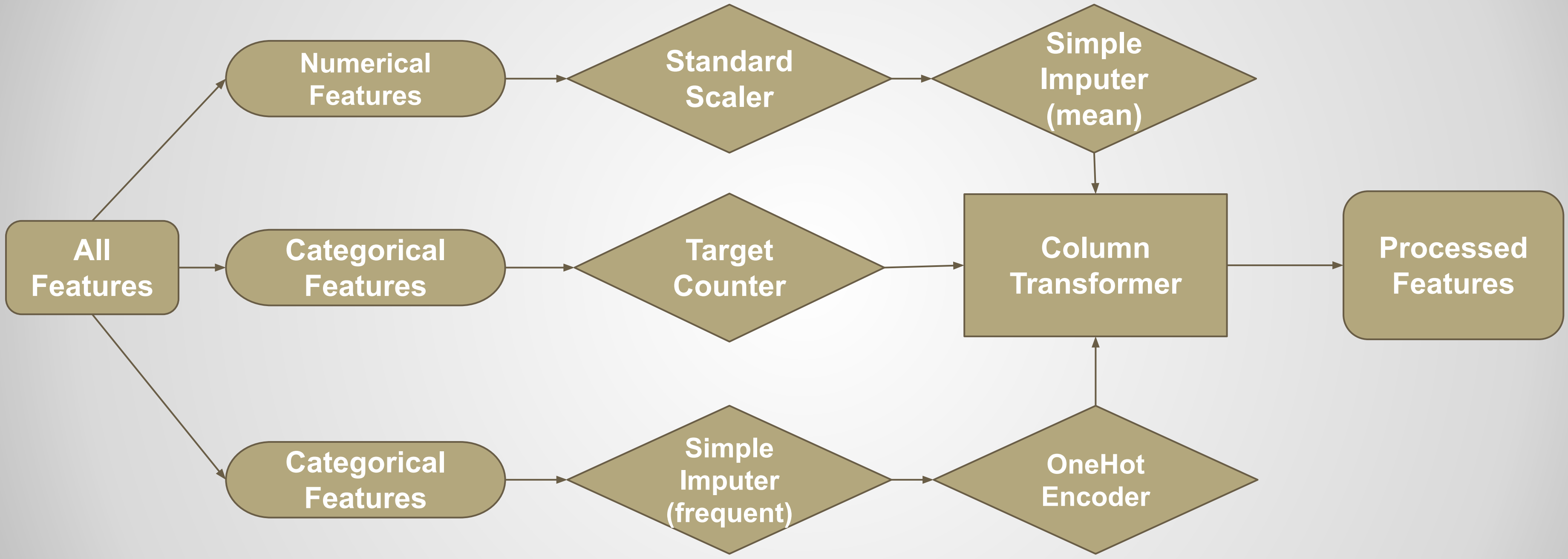
Numerical Data: credit limit, billing amount, payment amounts

Training/ Testing Data Points: 22500/ 7500

Problem Statement

What can we learn from using Machine Learning models utilized on previous credit card data to predict likelihood of future default credit payments?

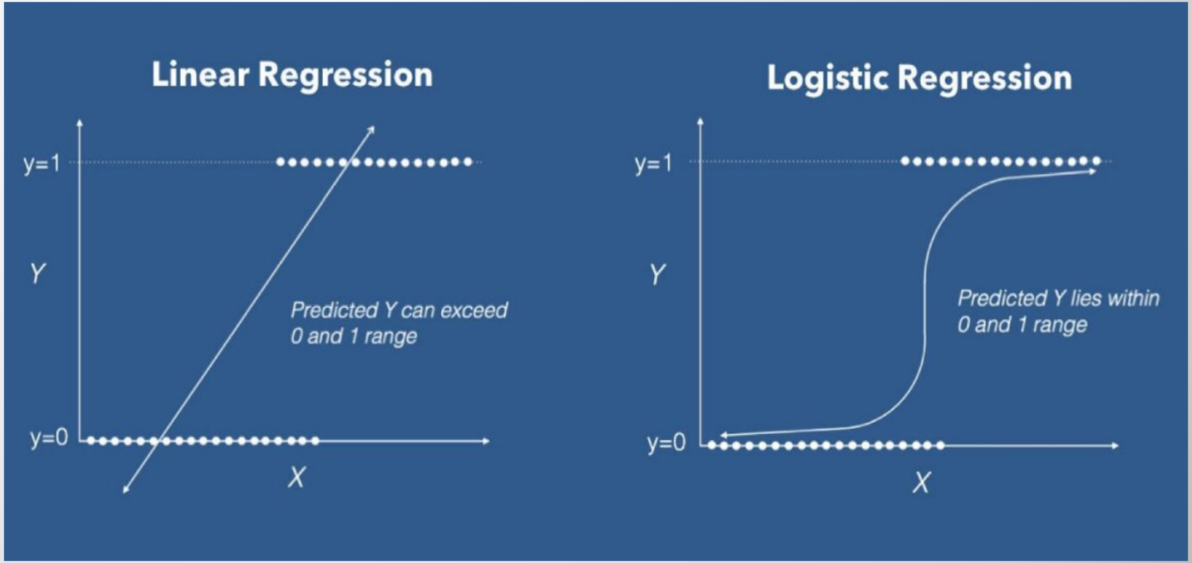
Data Processing Pipeline



Models

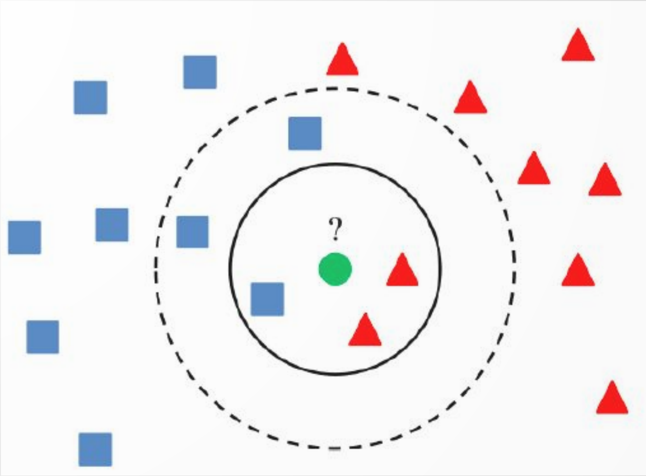
Logistic Regression

- Model prominently used for Binary Classification
- Sigmoid Function - range of score between 0 & 1
- Hyperparameters
  - C = 5
  - Penalty = l2
- **AUC 0.7621**



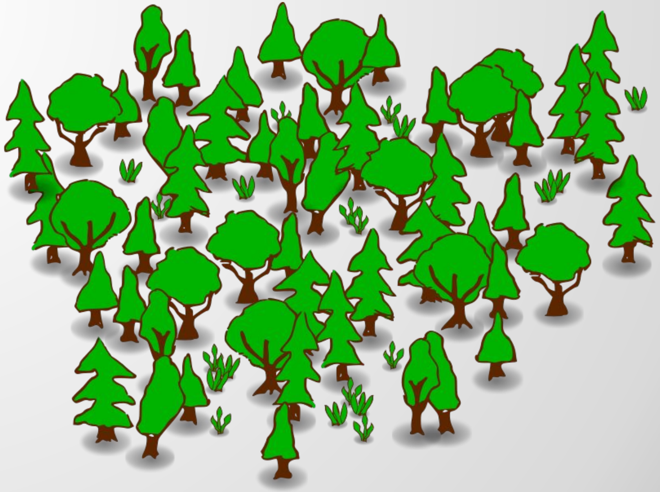
KNN Classifier

- K stand for the nearest neighbor
    - the key deciding factor in this algorithm
- Hyperparameters
    - N_neighbors =21
    - P (power parameter) =1
        - Manhattan_distance
    - Leaf_size =30
- **AUC 0.707**



Random Forest Classifier

- Predicts with multiple individual decision trees
- Reduces chances of overfitting
- Hyperparameter tuning **AUC 0.78998**
  - n_estimators = 600 (range 200-700)
  - max_depth = 11 (range 4-15)
- Grid search **AUC 0.78883**
  - n_estimators = 575
  - max_depth = 11



XGBoost

not
XGH !

Why: highly flexible and versatile, fast and accurate, being a good ensemble model to prototype your own ML projects

What: fits better structure / tabular datasets, recommended for regression and classification problems

It has become one of the most important models for Kaggle competitors willing to do well in online challenges

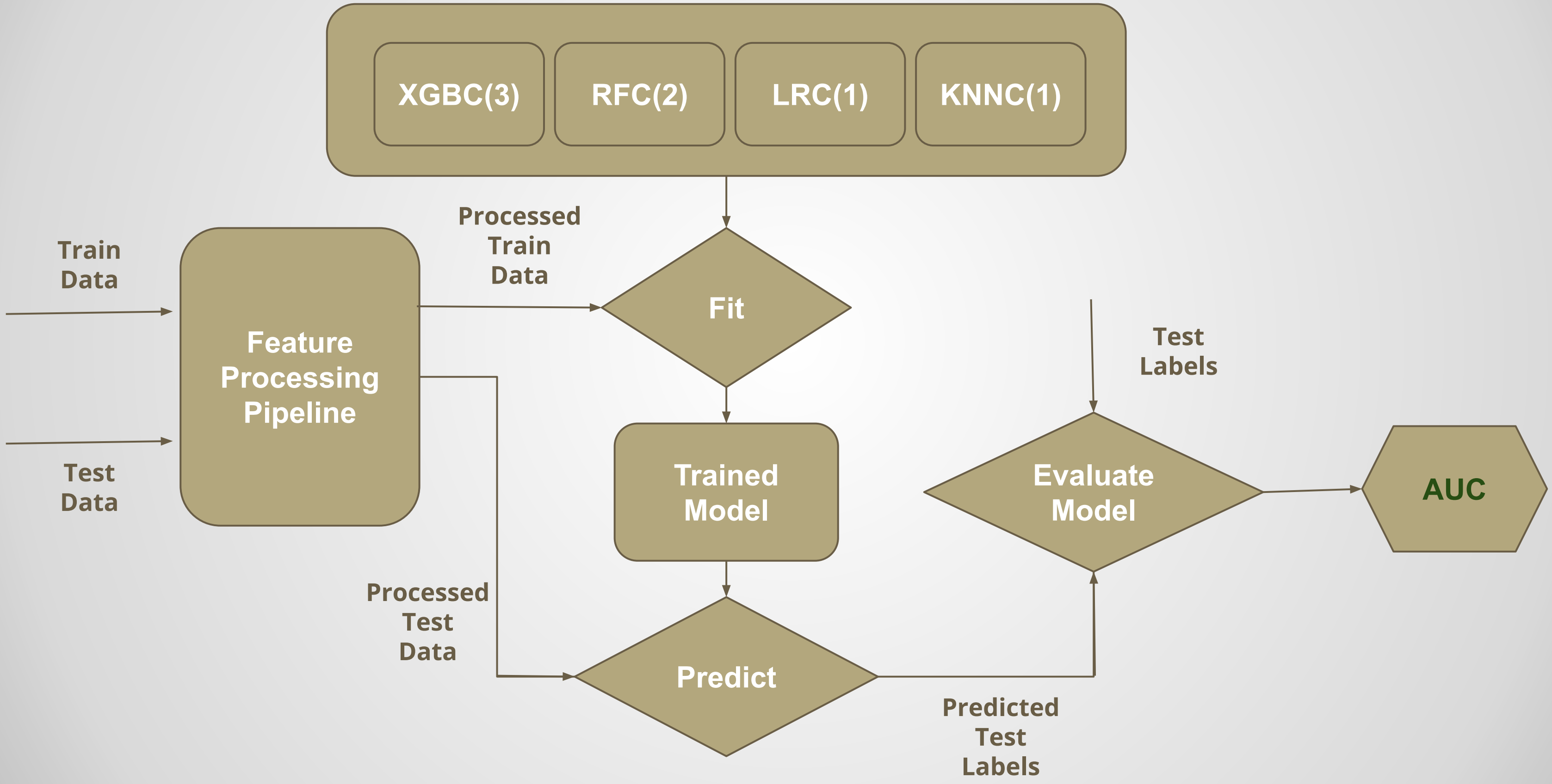
0.7971 AUC

CatBoost

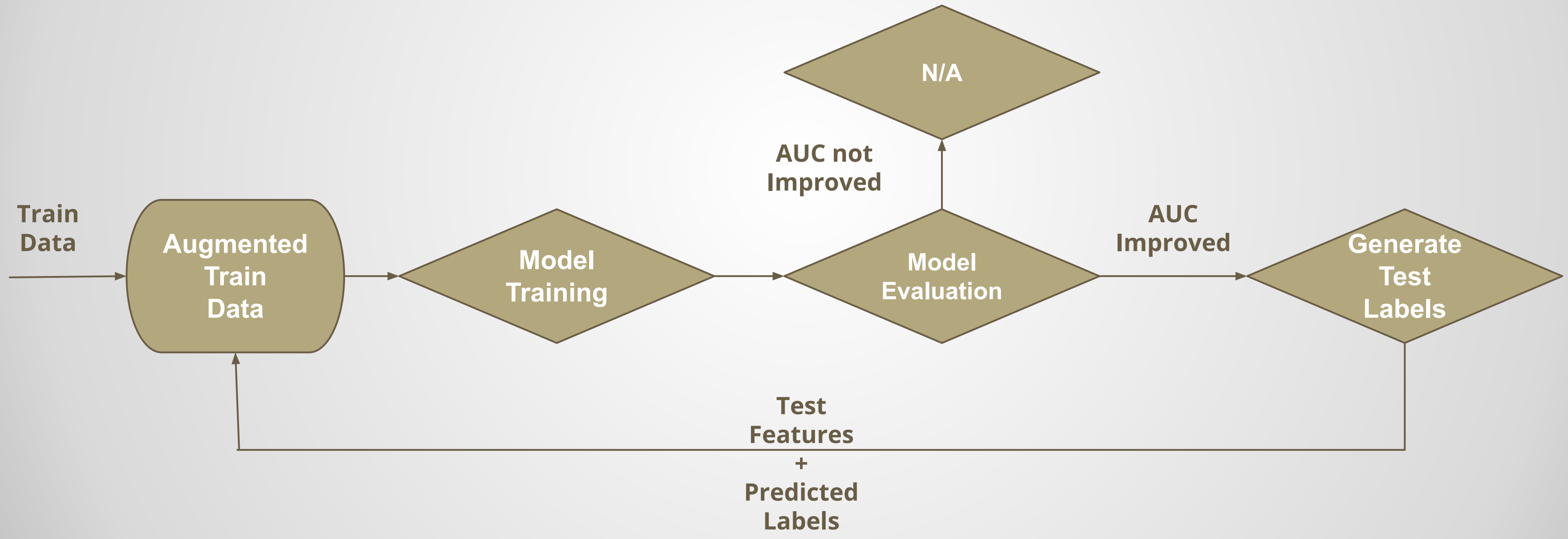
- Boosting method that targets categorical features
- Credit dataset has many categorical features:
  - Gender
  - Age
  - Marital status
  - Previous payment history
- Can tune parameters to adjust learning model

Our tuned model resulted in a 0.7944 AUC

VotingClassifier



Pseudo Labeling



Conclusions

1. While each model appears viable, most effective are Boosting methods, which we weigh more in VotingClassifier

2. Imbalanced dataset may give misleadingly accurate results

3. Different business objectives may result in different result evaluations
   a. Predicting defaults will prioritize avoiding false negatives
   b. Preventing defaults will want to focus on lowering false positives