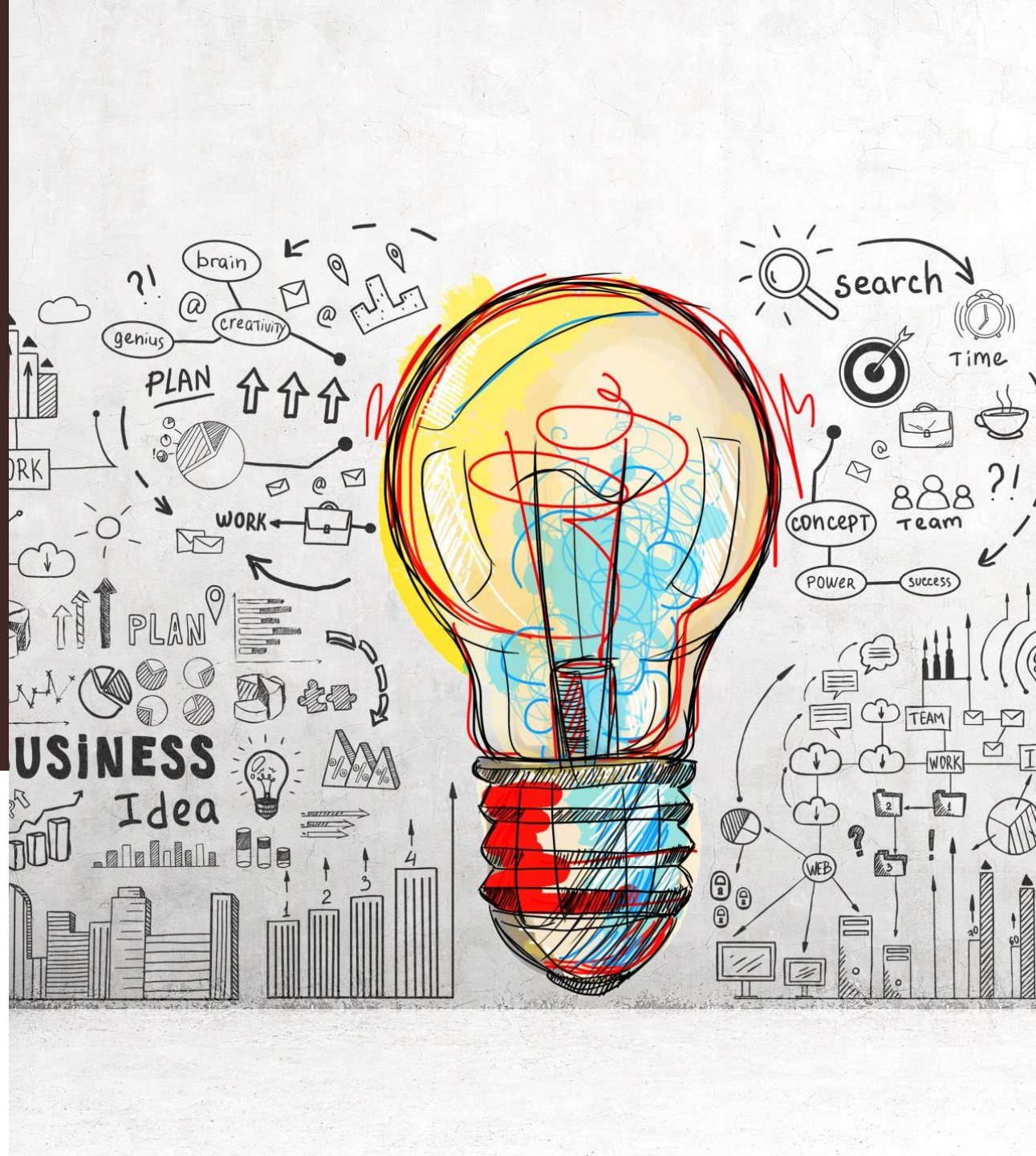
Applied Machine Learning Spring 2024 Group 9 Home Credit Default Risk



Paul Miller, Glen Colletti, Alex Bordanca



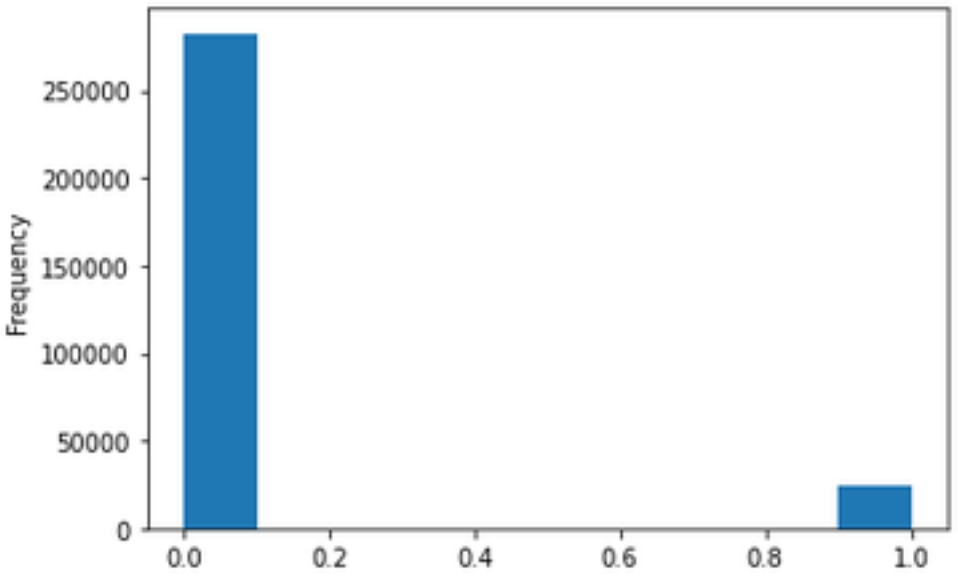
Project Description

- Home Credit Default risk
  - Predict from a variety of non-traditional credit metrics if a loan will be repaid
  - Kaggle competition for Home Credit company
    - Provide loans to those without traditional financial history
    - Mitigate risk
  - Preprocessing
  - Feature selection
  - Feature Engineering
  - Implement XG Boost, Logistic Regression, KNN classifier models

EDA

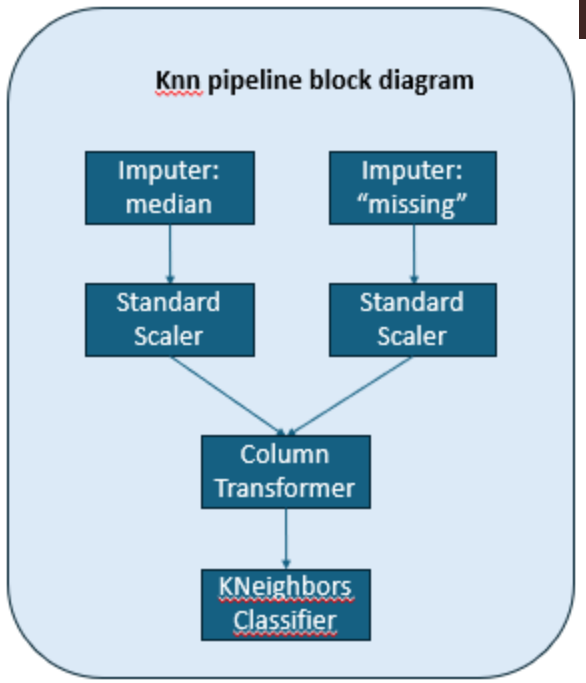
- 92% of target data is zero indicating the loan was repaid
  - 8% of loans were not repaid
    - Target feature is highly imbalanced, more noticeable with small sample size
- 121 features in data
  - 122 with 'TARGET'
- Identified the 10 most positively and 10 most negatively correlated features with 'TARGET'
- Null analysis, 50 features greater than 20% null

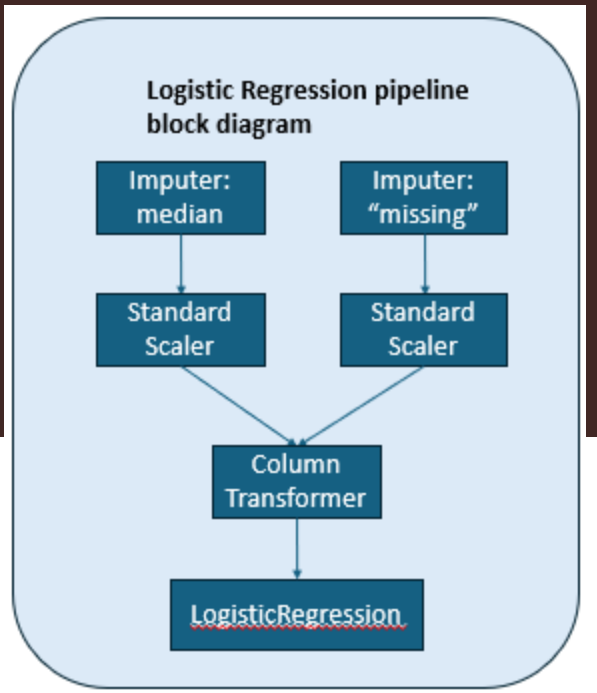
Frequency Plot 'TARGET' Feature

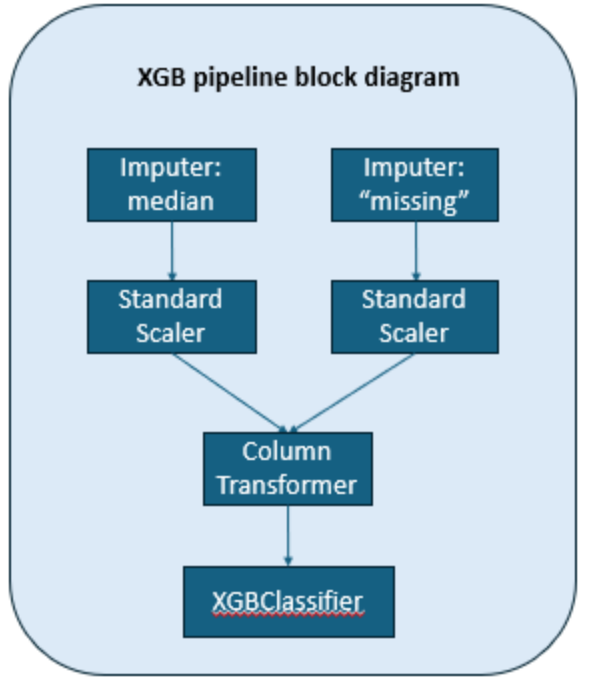


Modeling Pipelines

- Baseline pipelines were implemented for our three models
  - KNN
  - Logistic Regression
  - XGBoost
- Minimal preprocessing was accomplished to allow classifiers to run







Results

XGB performed best and was the only model with a non-zero F1 score meaning it was the only model with at least one True Positive

F1 is a blend of precision and recall.

Precision is defined as $$\frac{\text{True Positives}}{\text{True Positives + False Positives}}$$

Recall is defined as $$\frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

	exp_name	Train Acc	Valid Acc	Test Acc	Train F1	Valid F1	Test F1
0	knn_baseline	0.9234	0.9282	0.9133	0.0876	0.0000	0.058
1	XGB_baseline	0.9638	0.9300	0.9127	0.7007	0.0165	0.015
2	logreg_baseline	0.9216	0.9329	0.9160	0.0000	0.0000	0.000