ADVANCED ANALYTICS EDGE: PREDICTING CREDIT SCORE

Dalal Alramadhan

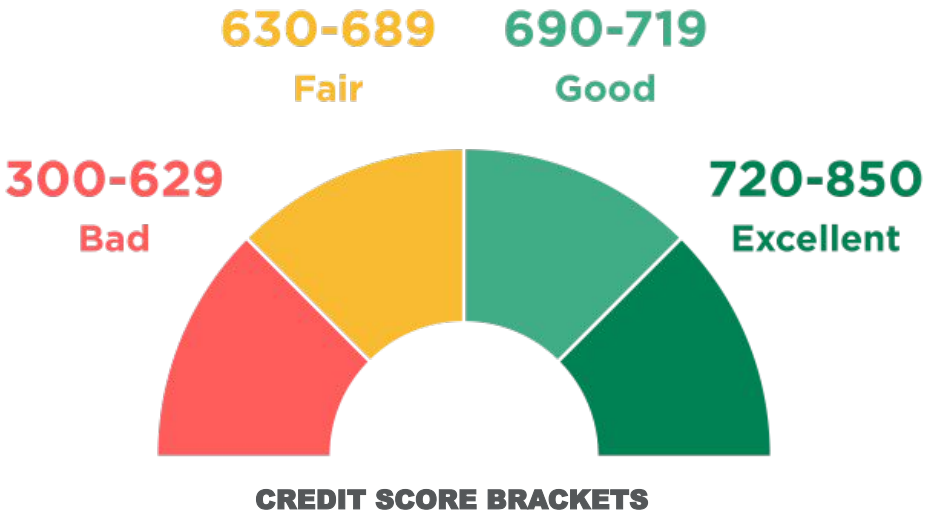
Brittany Nguyen

Nikita Singh

MIT SLOAN SCHOOL OF MANAGEMENT

MIT MANAGEMENT SLOAN SCHOOL

Introduction



Credit score numbers assess the likelihood of an individual or business to repay loans in a timely manner.

Factors influencing credit score: occupation, annual income, number of bank accounts/credit cards, etc.

MIT MANAGEMENT SLOAN SCHOOL

Objectives

PREDICT INDIVIDUAL'S CREDIT SCORE BRACKET

Predict "Poor", "Standard", "Good" using economic and loan status factors.

ASSESS VARIABLE IMPORTANCE

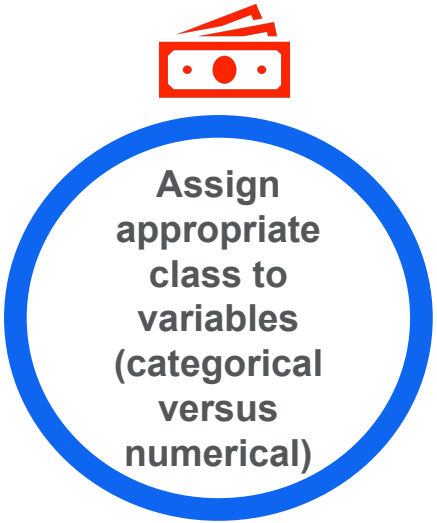
Assess which variables are most important in predicting credit scores.

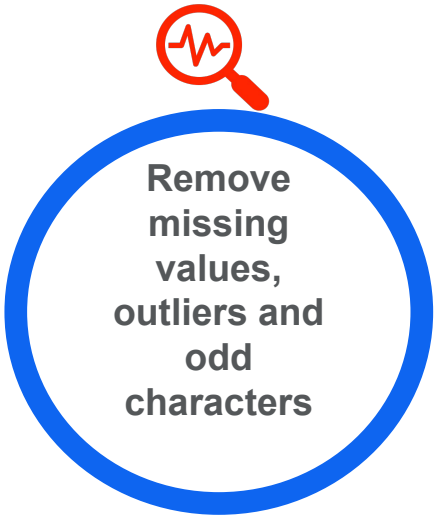
EVALUATE BEST PERFORMING MODELS

Evaluate which models result in the best performance metrics (AUC, accuracy)

Data Preprocessing

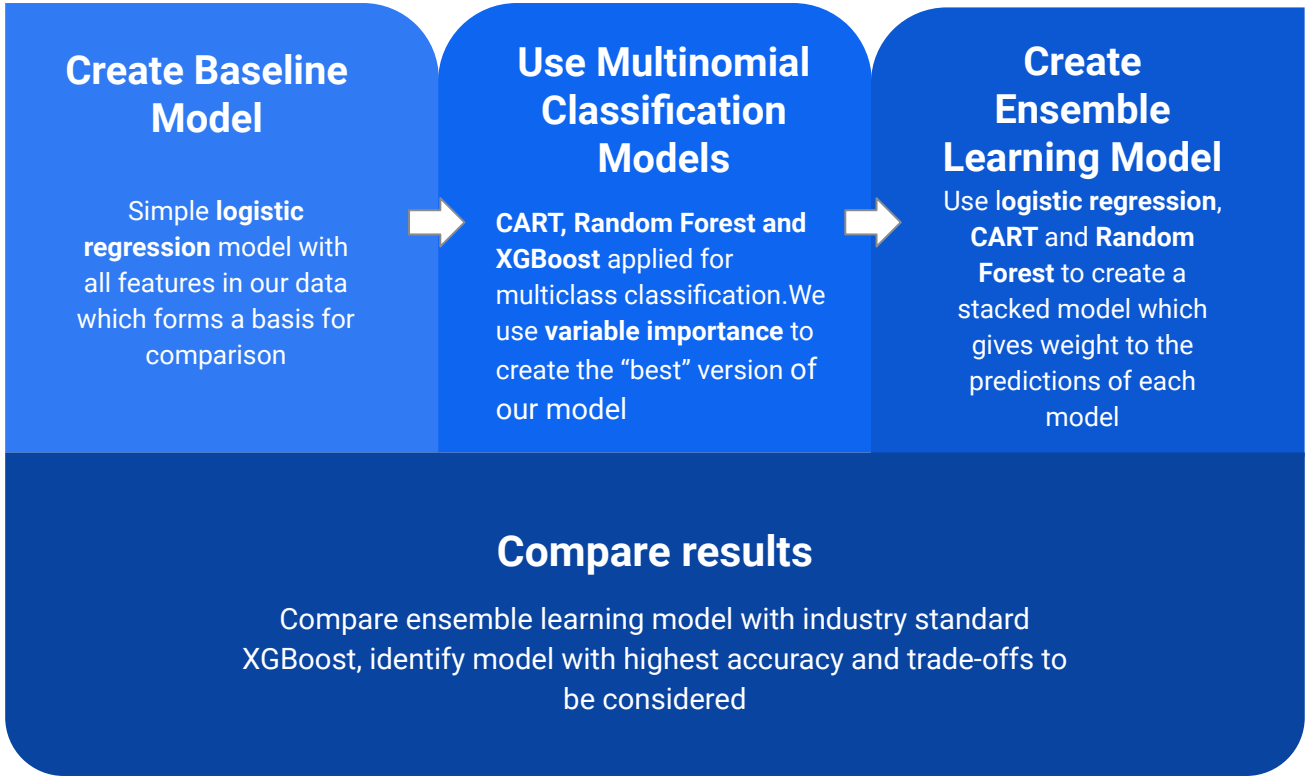
Data: **100,000 rows, 28 Independent Variables** including: “Month,” “Age,”, “Annual_Income”, “Interest_Rate” and “Num_Credit_Card”, **Dependent Variable:** Credit Score





MIT MANAGEMENT SLOAN SCHOOL

Methodology



MIT MANAGEMENT SLOAN SCHOOL

Implementation: CART

CART is a decision tree that explains how the target variable's values can be predicted based on other values

The challenges of the model: a small change in the dataset can make the tree structure unstable, causing variance

The benefits of the model: easy to understand, interpret, and visualize

Results
Accuracy: 62.47%
AUC: 0.680

Variable Importance:
Outstanding debt, interest rate, occupation, delay from due date, changed credit limit

Comparison with Baseline:
+1.96 percentage points in Accuracy
+0.002 increase in AUC

MIT MANAGEMENT SLOAN SCHOOL

Implementation: Random Forest

Random forest is an ensemble learning method for classification that constructs a multitude of decision trees

The challenges of the model: long computational power and time as it combines a lot of decision trees to determine the class

The benefits of the model: reduces overfitting in decision trees and helps to improve the accuracy (“power in numbers”)

Results
Accuracy: 80.24%
AUC: 0.818

Variable Importance:
Outstanding debt, interest rate, occupation, delay from due date, changed credit limit

Comparison with Baseline:
+19.73 percentage points in Accuracy
+0.136 increase AUC

MIT MANAGEMENT SLOAN SCHOOL

Implementation: XGBoost

XGBoost is commonly considered as the **industry standard** in regression, classification and ranking problems

Applying this model to our problem was **challenging** as multiclass labels need to be processed further in a specific format

With cross-validation, the **processing time** increases. However, the computational load is awarded with **high accuracy** and **AUC**

Results
Accuracy: 90.36%
AUC: 0.9496

Comparison with Baseline:
+29.85 percentage points in Accuracy
+0.2673 increase AUC

MIT MANAGEMENT SLOAN SCHOOL

Implementation: Ensemble Model

An interesting approach that we implemented was a **stacked model** consisting of logistic regression, CART and Random Forest

We wished to see if gathering these models in an ensemble would (a) **improve upon performance of individual models** and (b) compare results with XGBoost

While the stacked model improved upon the accuracy of baseline and CART, it could not outperform RF and XGBoost

Results
Accuracy: 79.49%
AUC: 0.7950

Comparison with Baseline:
+18.98 percentage points in Accuracy
+0.1065 increase AUC

Comparison with XGBoost:
-10.87 percentage points in Accuracy
-0.1546 decrease AUC

MIT MANAGEMENT SLOAN SCHOOL

Results and Model Comparison

	Baseline	CART	Random Forest	XGBoost	Ensemble Learning Model
Accuracy	**60.51%**	**62.47%** With feature selection: **63.4%**	**80.24%** With feature selection: **79.70%**	**90.36%**	**79.49%**
AUC	**0.682**	**0.680** With feature selection: **0.6885**	**0.818** With feature selection: **0.8040**	**0.949**	**0.795**
Interpretability	Logistic regression: not very interpretable	**Highly interpretable**	Not interpretable	Highly uninterpretable	Not interpretable
Complexity	Low complexity, no cross-validation	**Low complexity, even with cross-validation**	High complexity	Very high complexity and computational cost	Somewhat complex

MIT MANAGEMENT SLOAN SCHOOL

Future Scope

Create thresholds for XGBoost classification

Currently, our XGBoost model predicts label based on the class which has the highest probability

However, a more sophisticated approach would be to create thresholds for each class and predict on that basis

Understand in-depth why credit score is low

Variable importance gives us a good idea about which features contribute to prediction of credit score bracket

But we can further take a prescriptive analytical view in understanding why an individual's credit score is low and what can be done to improve it

Improved credit score categorization system

While the categorization we have utilized is industry standard, it is not intuitive and interpretable

A more systematic bracketing system will be useful

MIT MANAGEMENT SLOAN SCHOOL