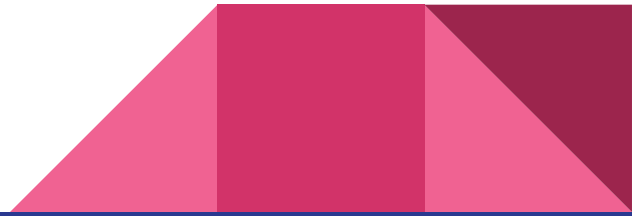
Home Credit Default Risk Prediction using Machine Learning methods

Meghana Kotrakona

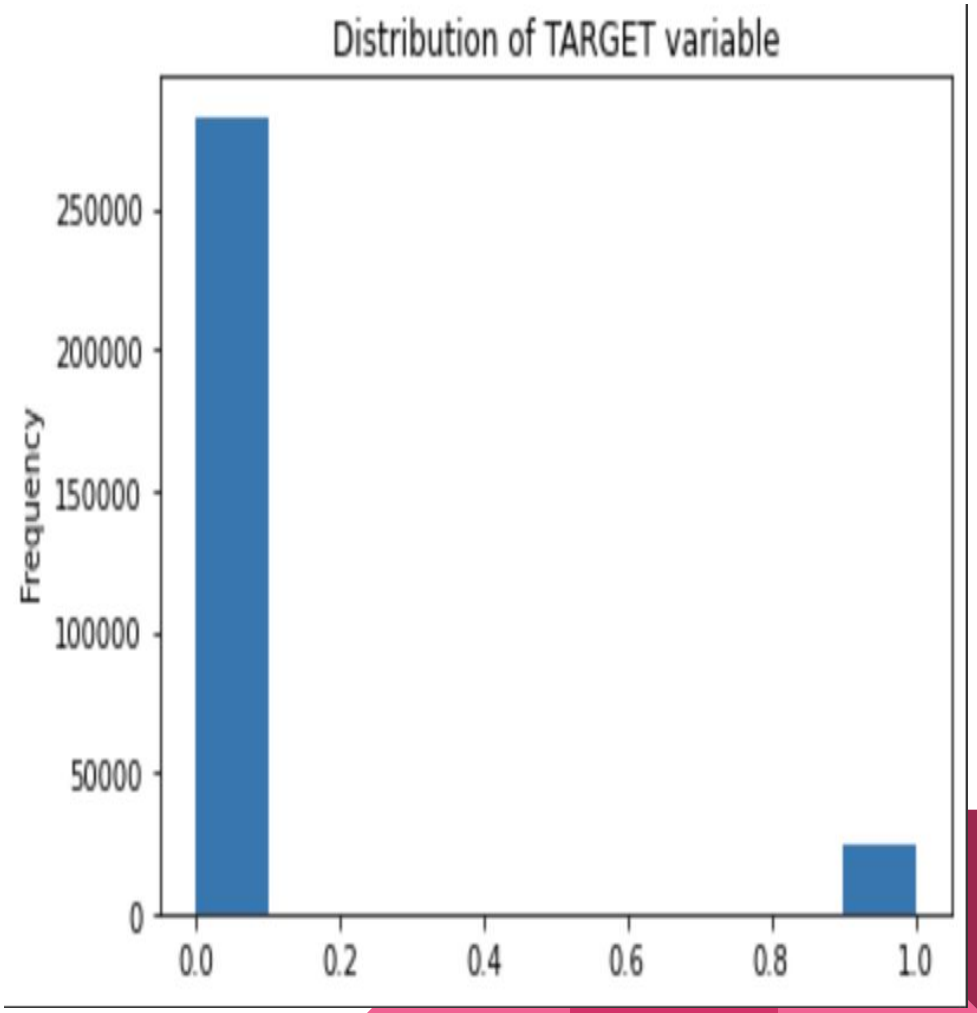
Business Objective

- Home Credit is a lending company providing loans to their customers.
- The problem faced by Home Credit group is the loss of money issued to defaulters, who are not able to repay. This impacts the company's financials.
- The solution attempted here is to build models to identify mortgage defaulters and help the company reduce the losses from these defaulted loans.
- This helps bank to provide loans to clients capable of repayment and saves lot of money and thereby become more profitable.



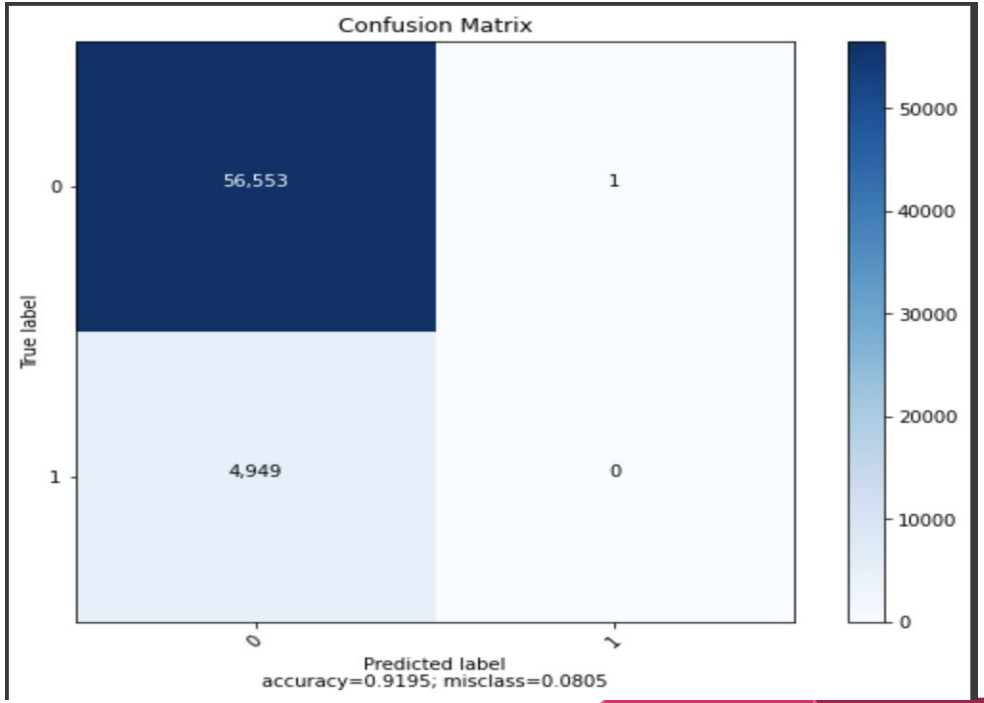
Data overview

- The data contains 307511 loan applications and 122 features with information about each loan application at Home Credit.
- The target variable defines whether the loan was repaid or not.
- The target variable is imbalanced with the majority of applicants has the target equals to zero.



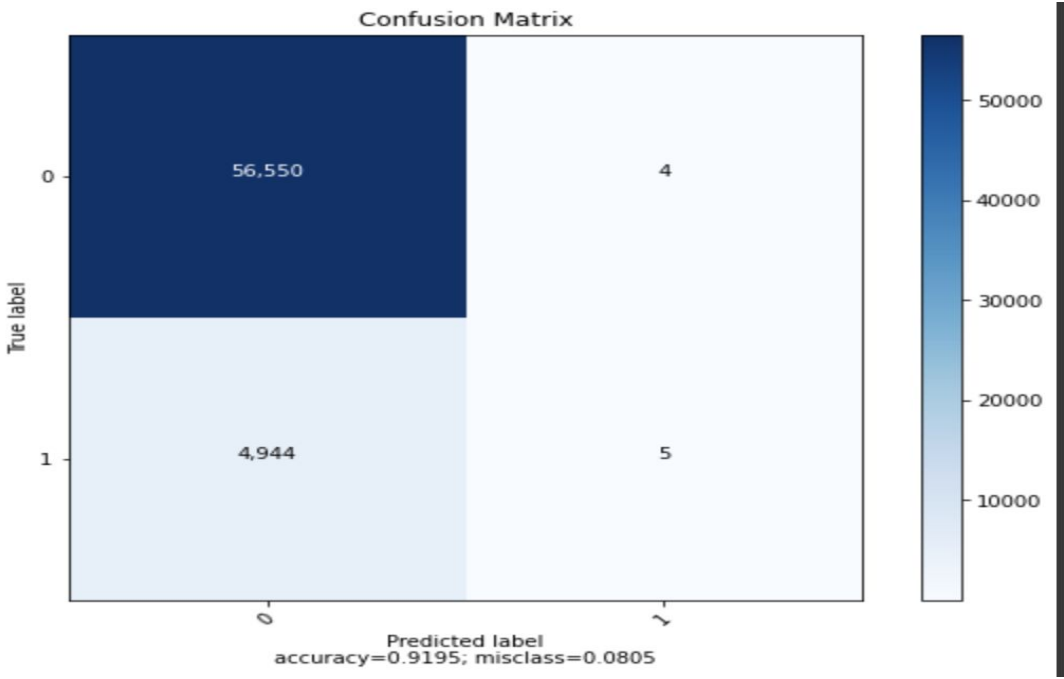
Baseline Model

- Using all features to fit a logistic regression model.
- Out of 4949 loans that are default this model hasn’t anything on default loan accurately.
- So both precision and Recall are zero.



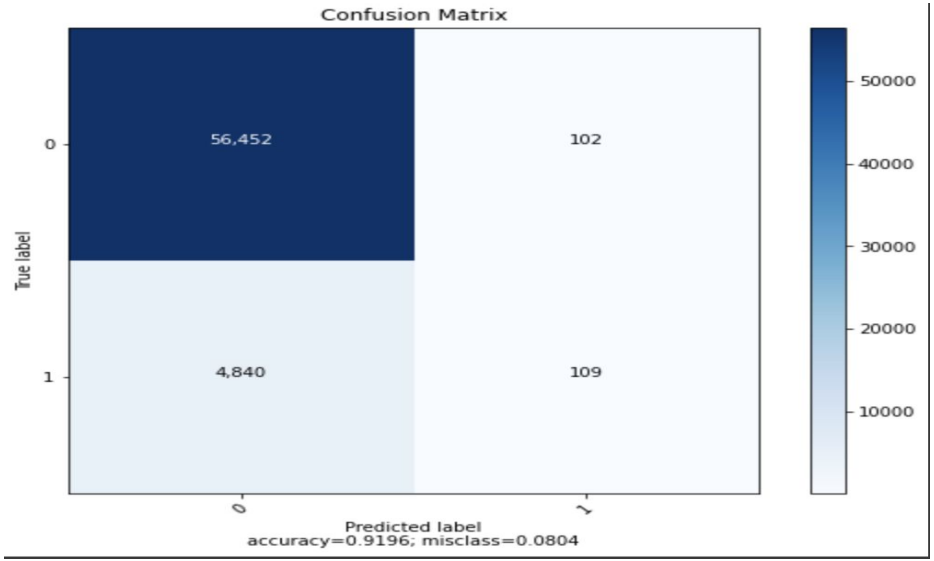
Random Forest Model

- Out of 4949 that are default, this model correctly predicted 5 as defaulters. So Recall is quite low.
- 4 applicants are wrongly classified as defaulters. This yields Precision score of 56%.



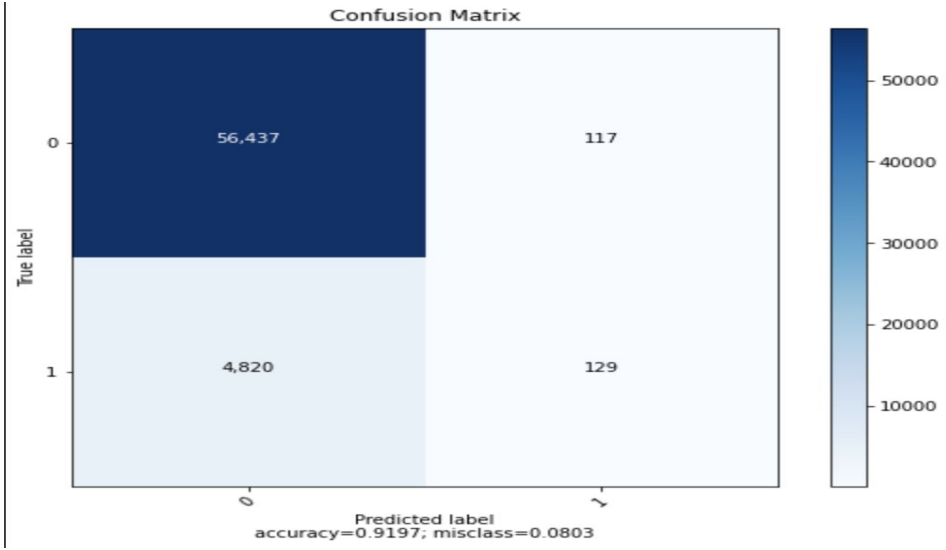
Gradient Boosting Model

- Out of 4949 that are default, this model correctly predicted 109 as defaulters. Recall is 2%.
- 102 applicants are wrongly classified as defaulters. This yields Precision score of 52%.
- Recall score is improved by 2% from Baseline.



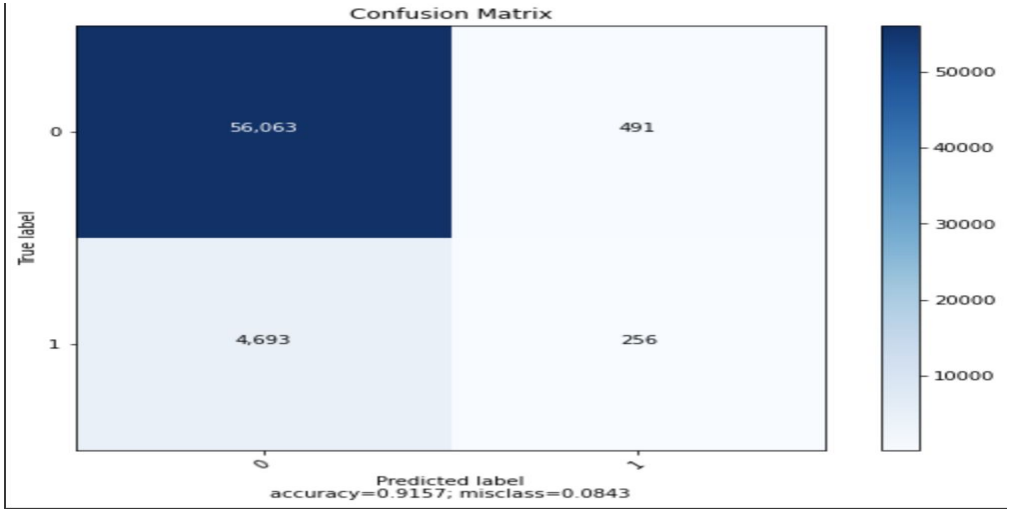
Gradient Boost Model including new variables

- Out of 4820 that are default, this model correctly predicted 129 as defaulters. So Recall is 3%.
- 117 applicants are wrongly classified as defaulters. This yields Precision score of 52%.
- Recall score is improved by 3% from Baseline.

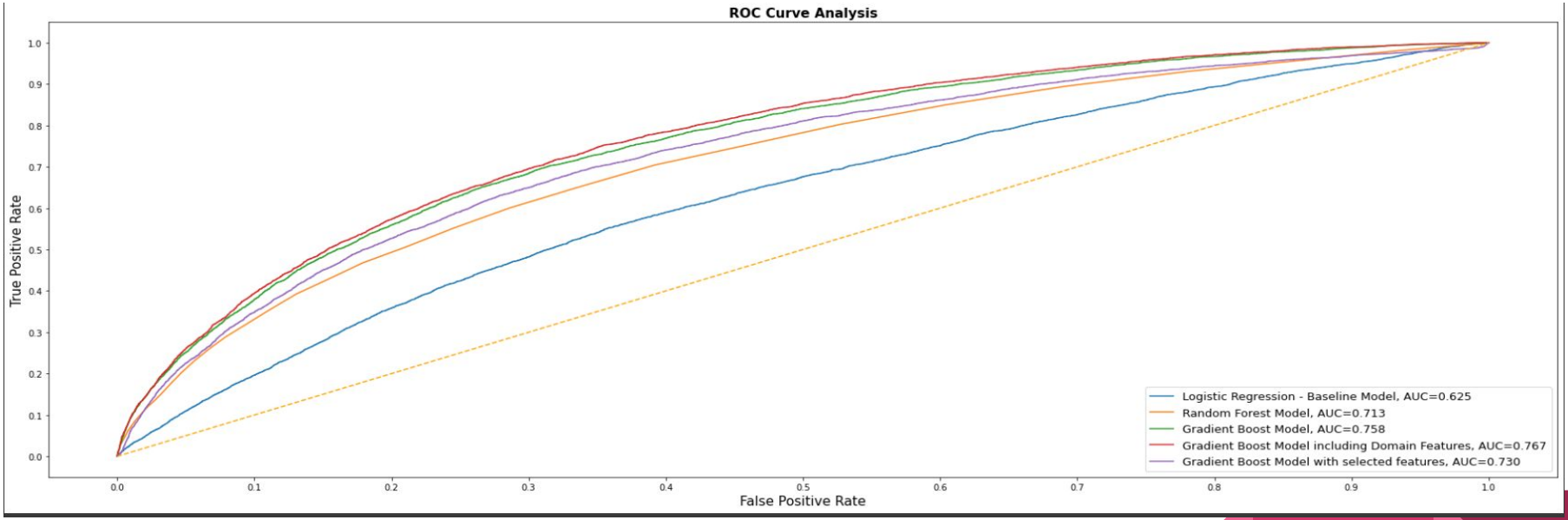


Gradient Boost Model using selected features

- Out of 4820 that are default, this model correctly predicted 256 as defaulters. So Recall is 5%.
- 491 applicants are wrongly classified as defaulters. This yields Precision score of 63%.
- Recall score is improved by 5% from Baseline.



Compare Models performance with ROC_AUC curves



Results

	Logistic Regression	Random Forest	Gradient Boost Model	Gradient Boost with new features	Gradient Boost with selected features
Accuracy	0.92	0.92	0.92	0.92	0.92
Precision	0	0.56	0.52	0.52	0.34
Recall	0	0	0.02	0.03	**0.05**
F1-score	0	0	0.04	0.05	**0.09**
ROC_AUC	0.625	0.713	0.758	**0.767**	0.730

Summary

- Gradient Boosting Model have the best results which gave ROC_AUC of 0.767, 5% Recall score, 9% F1 Score, which in turn saved the business $126M.
- In the future, models can be further improved by dealing more carefully with missing values, implementing better strategies for selecting features, getting more information about the clients by which the Credit company can save more money.