Home Credit Default Risk Prediction using Machine Learning methods

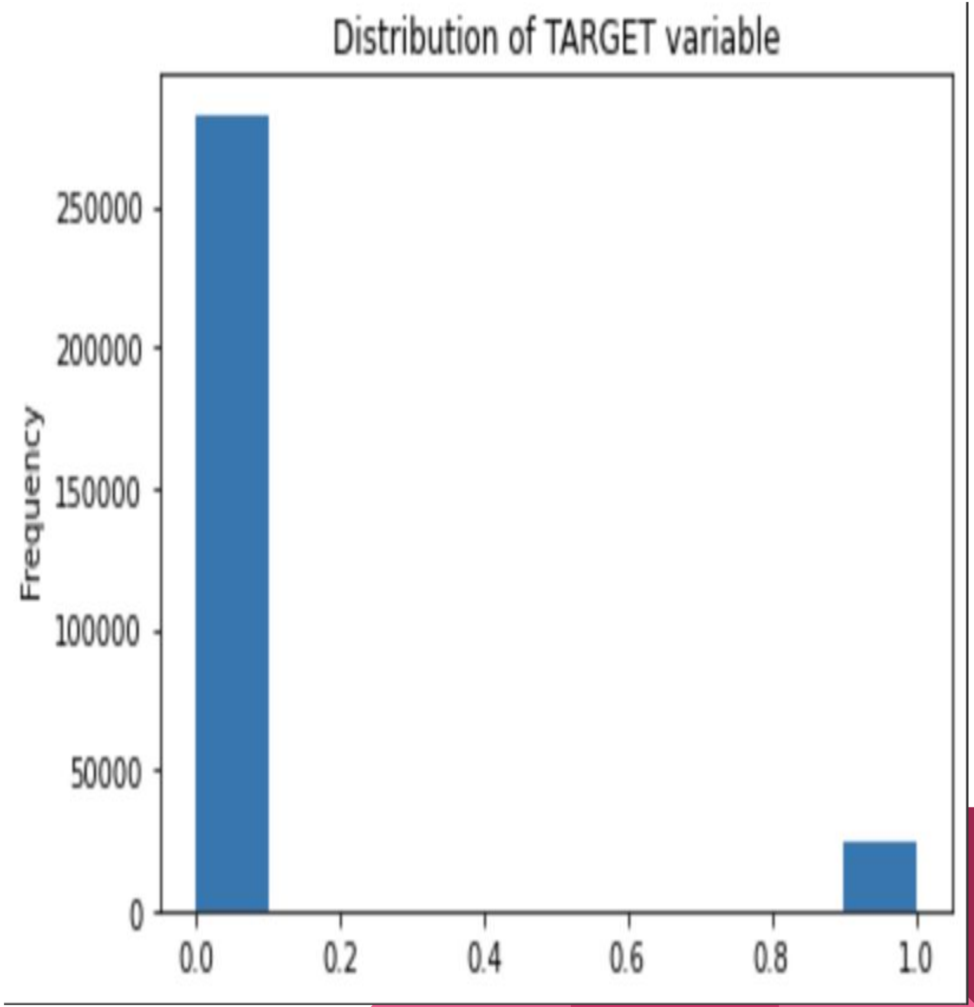
Meghana Kotrakona

Overview of Home Credit Default Risk Prediction

- Home Credit is a financial institution focussing to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience.
- The dataset exposed by Home Credit involves historical loan related details offered to their clients.
- Using this data, I intend to use several models to identify whether an applicant is capable of paying a loan or not.
- Doing so will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.

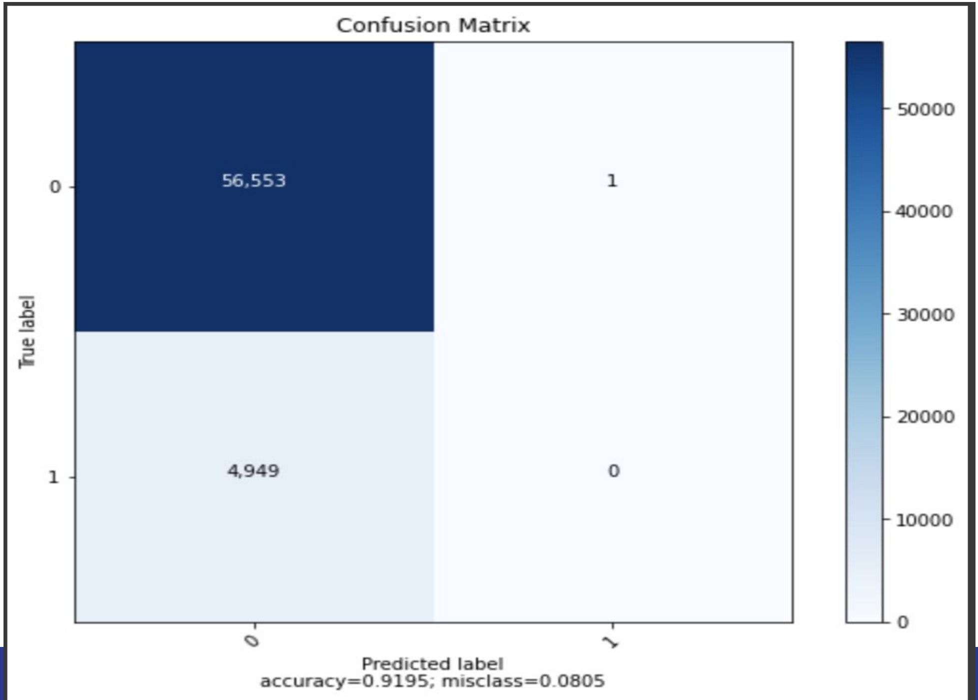
Data overview

- Data is provided by Home Credit.
- 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 repayed or not.
- The target variable is imbalanced with the majority of applicants has the target equals to zero



Baseline Model

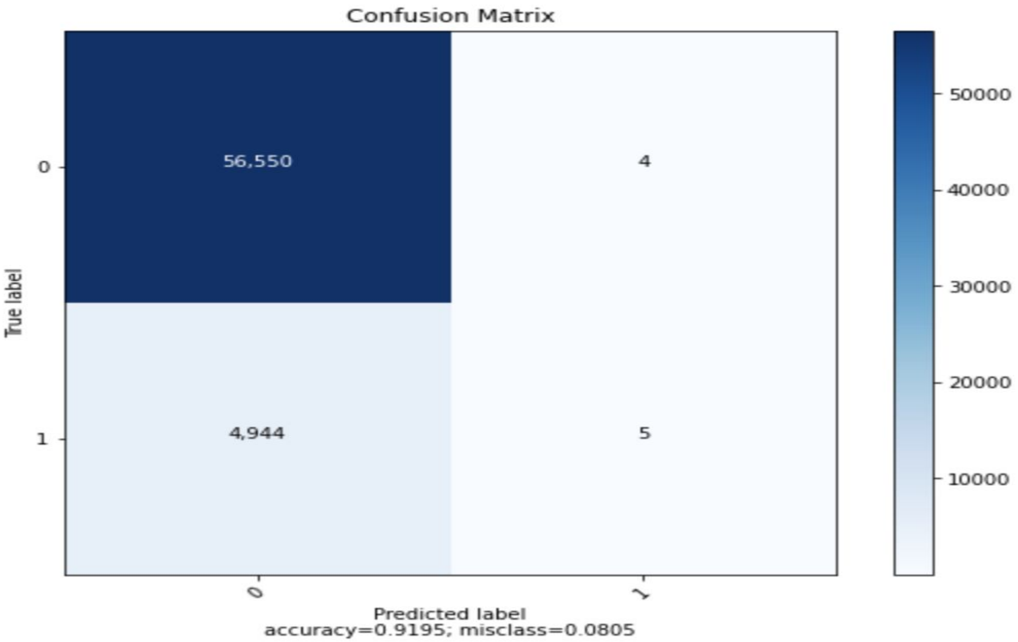
- Using all features to fit a logistic regression model.
- No instances correctly predicted in non-defaulter class. The performance is poor.



	precision	recall	f1-score	support
0	0.92	1.00	0.96	56554
1	0.00	0.00	0.00	4949
accuracy			0.92	61503
macro avg	0.46	0.50	0.48	61503
weighted avg	0.85	0.92	0.88	61503

Random Forest Model

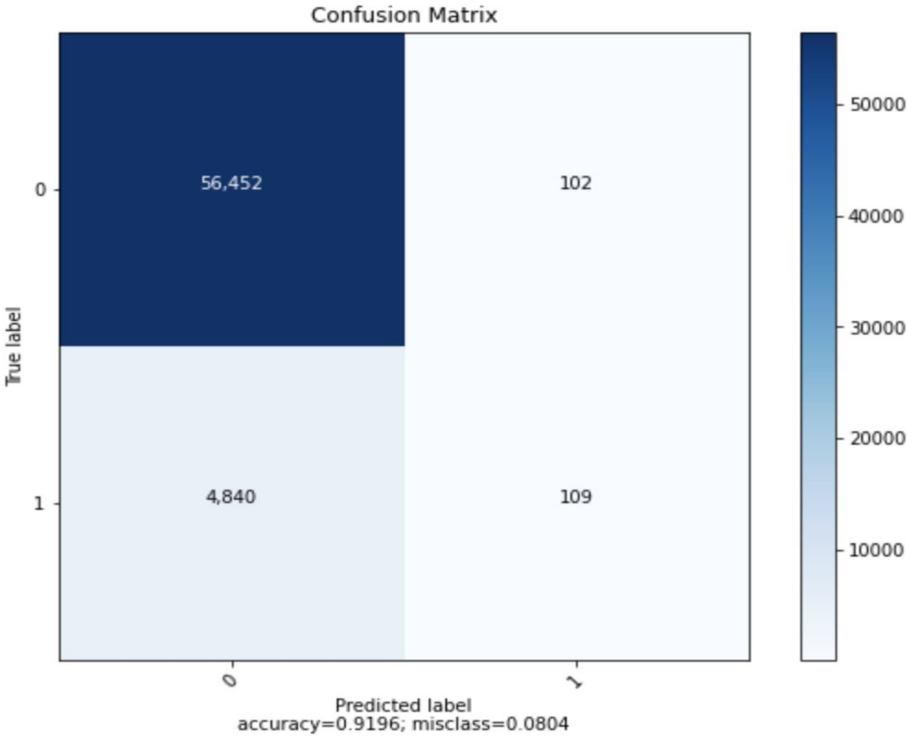
- Random Forest Model using 100 estimators
- K-fold cross validation to evaluate performance on training set.



	precision	recall	f1-score	support
0	0.92	1.00	0.96	56554
1	0.56	0.00	0.00	4949
accuracy			0.92	61503
macro avg	0.74	0.50	0.48	61503
weighted avg	0.89	0.92	0.88	61503

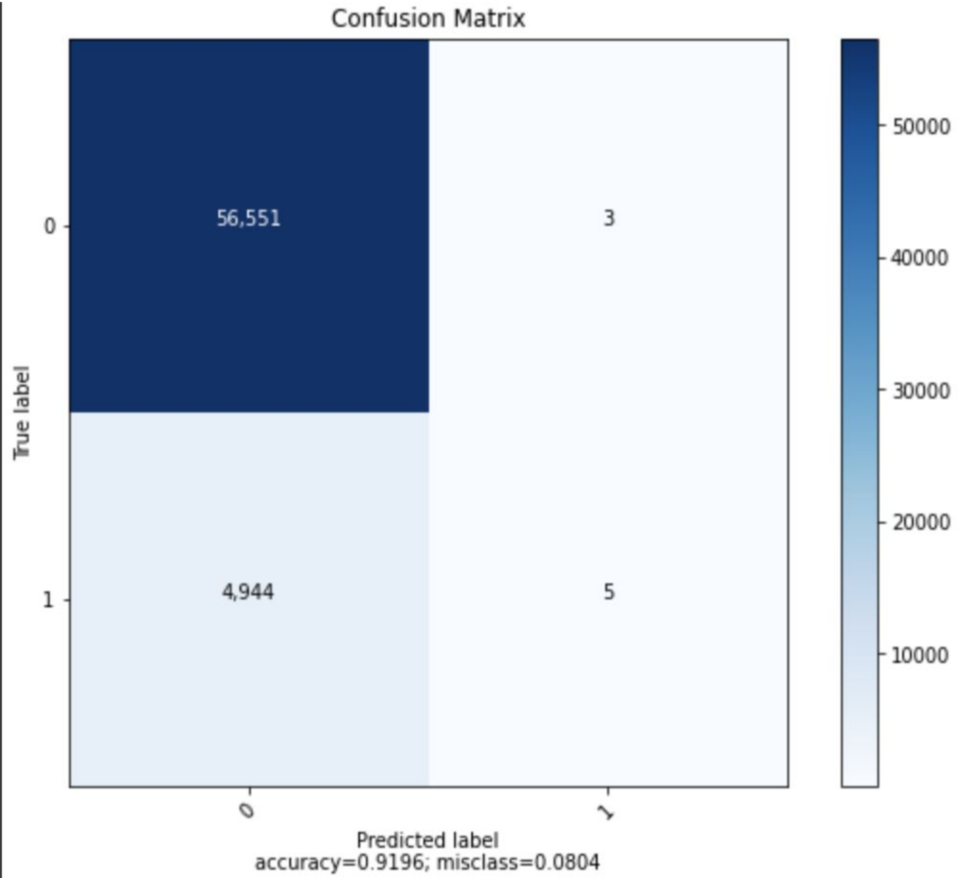
Gradient Boost Model

Built making 500 iterations, 2 max depth



	precision	recall	f1-score	support
0	0.92	1.00	0.96	56554
1	0.52	0.02	0.04	4949
accuracy			0.92	61503
macro avg	0.72	0.51	0.50	61503
weighted avg	0.89	0.92	0.88	61503

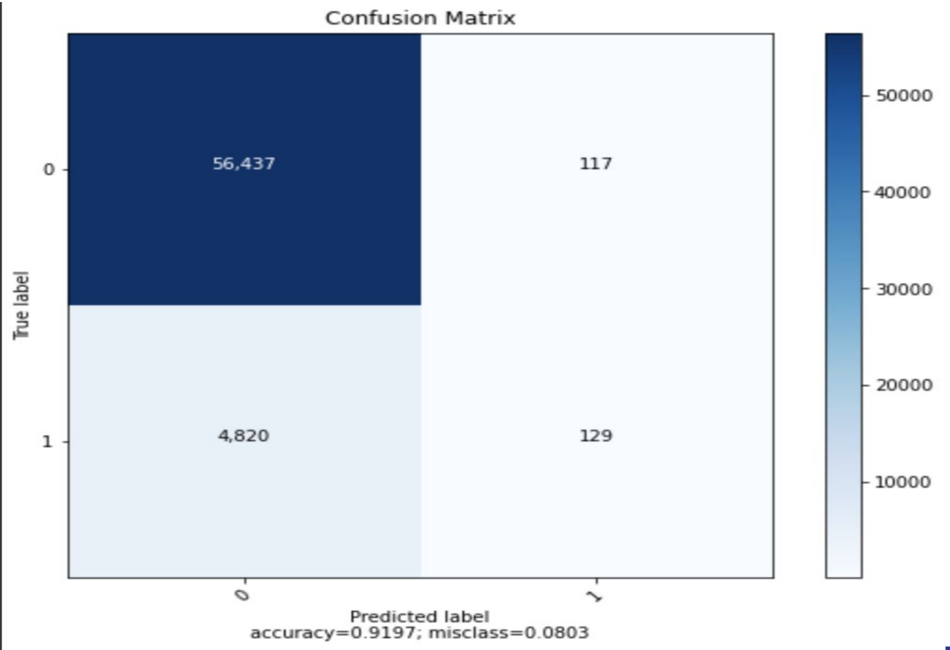
Random Forest Mode including new variables



	precision	recall	f1-score	support
0	0.92	1.00	0.96	56554
1	0.62	0.00	0.00	4949
accuracy			0.92	61503
macro avg	0.77	0.50	0.48	61503
weighted avg	0.90	0.92	0.88	61503

Gradient Boost Model including new variables

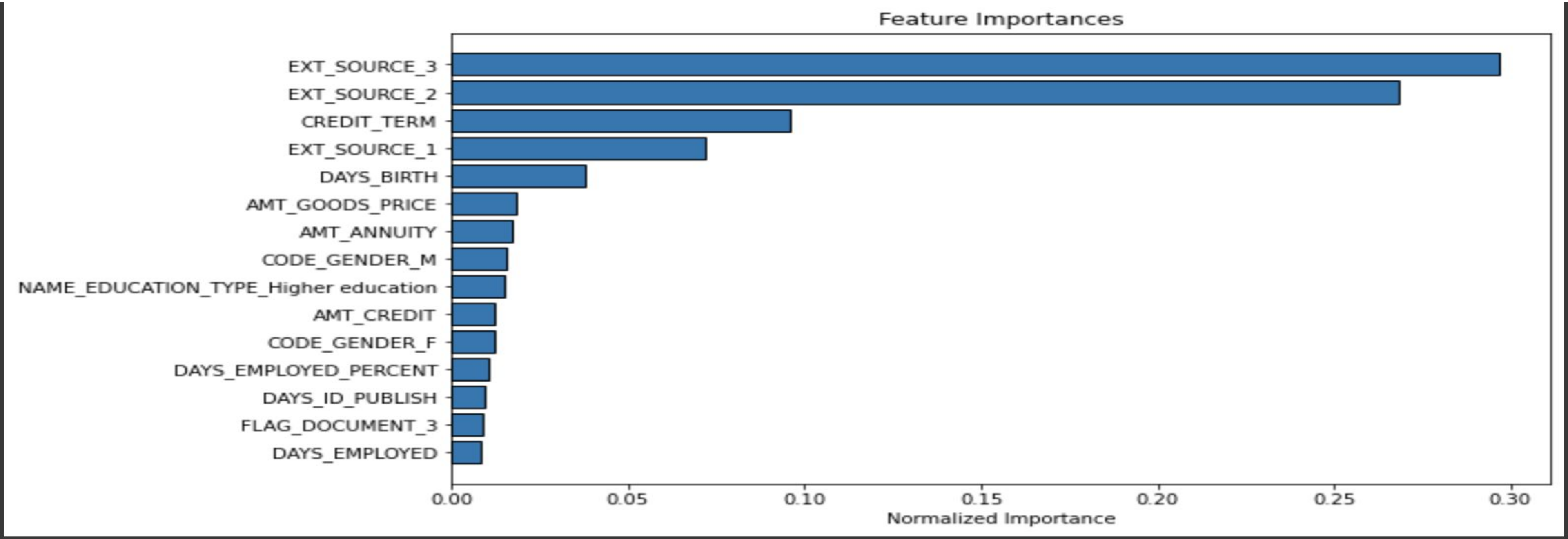
Recall and F1 score are improved



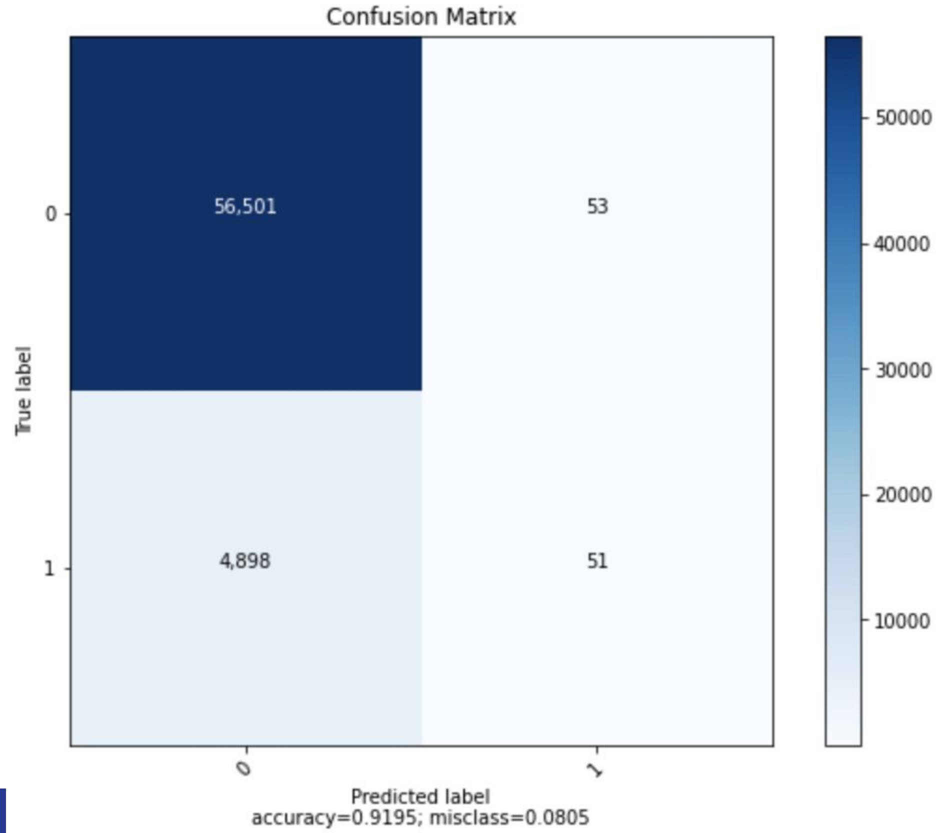
	precision	recall	f1-score	support
0	0.92	1.00	0.96	56554
1	0.52	0.03	0.05	4949
accuracy			0.92	61503
macro avg	0.72	0.51	0.50	61503
weighted avg	0.89	0.92	0.88	61503

Feature selection

Feature Importances from previous Gradient Boost Model



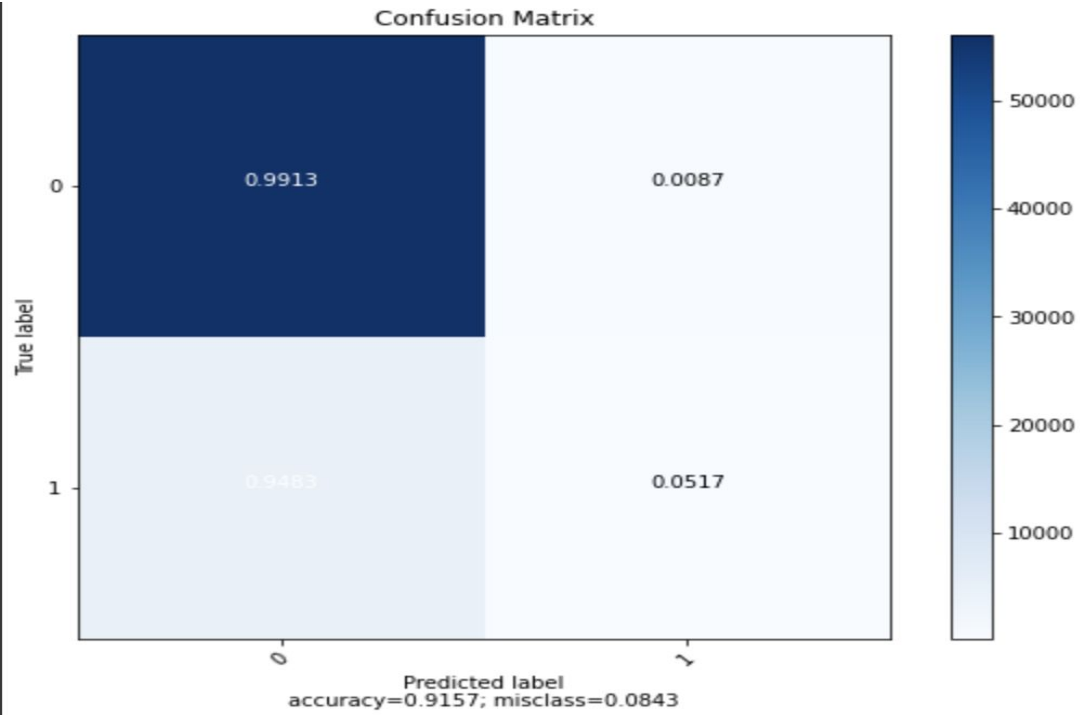
Random Forest Model using selected features



	precision	recall	f1-score	support
0	0.92	1.00	0.96	56554
1	0.49	0.01	0.02	4949
accuracy			0.92	61503
macro avg	0.71	0.50	0.49	61503
weighted avg	0.89	0.92	0.88	61503

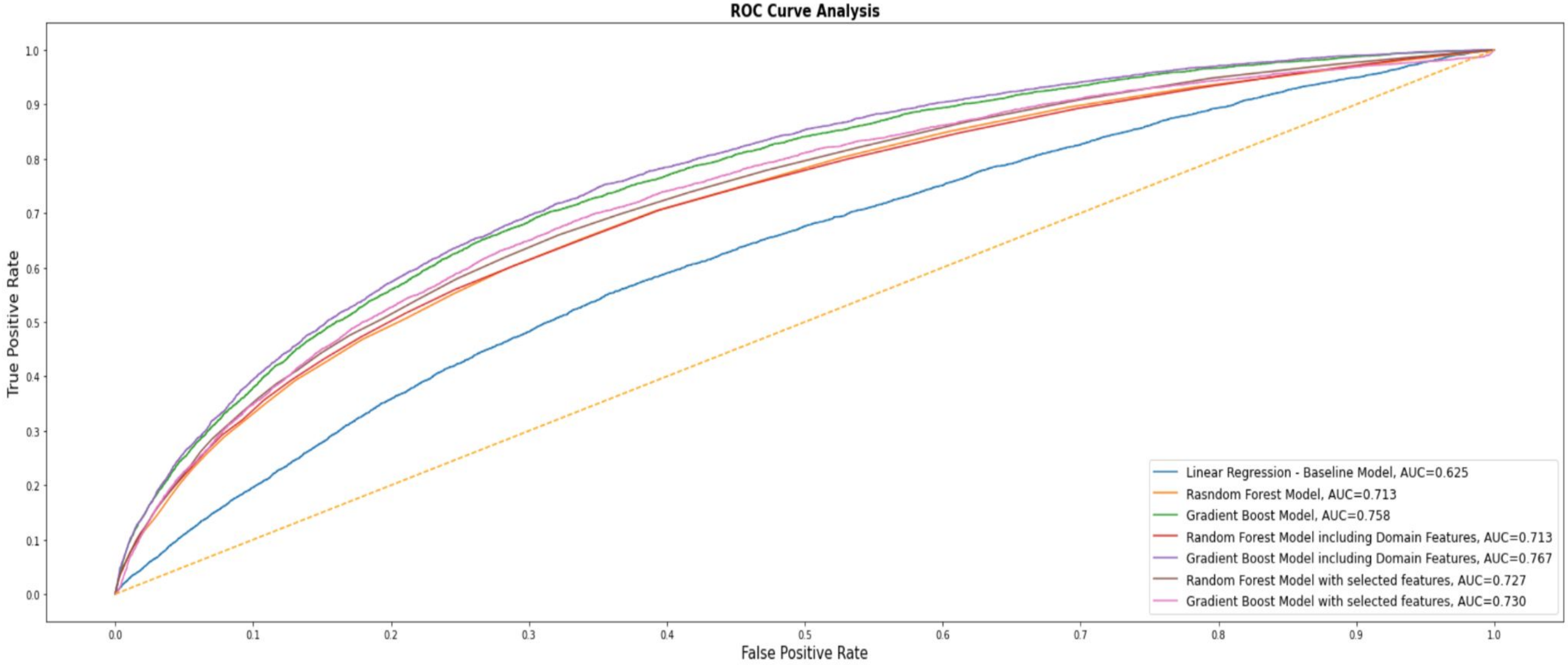
Gradient Boost Model using selected features

Better Recall and F1 score



	precision	recall	f1-score	support
0	0.92	0.99	0.96	56554
1	0.34	0.05	0.09	4949
accuracy			0.92	61503
macro avg	0.63	0.52	0.52	61503
weighted avg	0.88	0.92	0.89	61503

Compare Models performance with ROC_AUC curves



Results

	Logistic Regression	Random Forest	Gradient Boost Model	Random Forest with new features	Gradient Boost with new features	Random Forest with selected features	Gradient Boost with selected features
Accuracy	0.92	0.92	0.92	0.92	0.92	0.92	0.92
Precision	0	0.56	0.52	**0.62**	0.52	0.49	0.34
Recall	0	0	0.02	0	0.03	0.01	**0.05**
F1-score	0	0	0.04	0	0.05	0.02	**0.09**
ROC_AUC	0.625	0.713	0.758	0.713	**0.767**	0.727	0.730

Summary

- Gradient Boosting Model have the best results.
- Gradient Boosting Model with selected features has dominant performance in Recall, F1 score. Gradient Boosting Model with new features also has good permanence in ROC_AUC.
- In the future, models can be further improved by dealing more carefully with missing values, implementing better strategies for selecting features and tuning hyperparameters more precisely.