CAPSTONE 1: LOAN ANALYSIS

Vy Nguyen

Challenges

- Small and medium enterprises (SMEs) are one of the strongest drivers of economic development, innovation and employment.
- The access to finance is recognized as the greatest obstacle to the growth of SMEs.
- Lenders who can address this matter have great growing potential.
- Challenges: how to manage loan risk efficiently and make accurate decisions.

Goal and Data

- **Goal:** Using Machine Learning models to predict whether a company is able to make payments on time.
- **Data**: The dataset is taken from the South African digital lender .

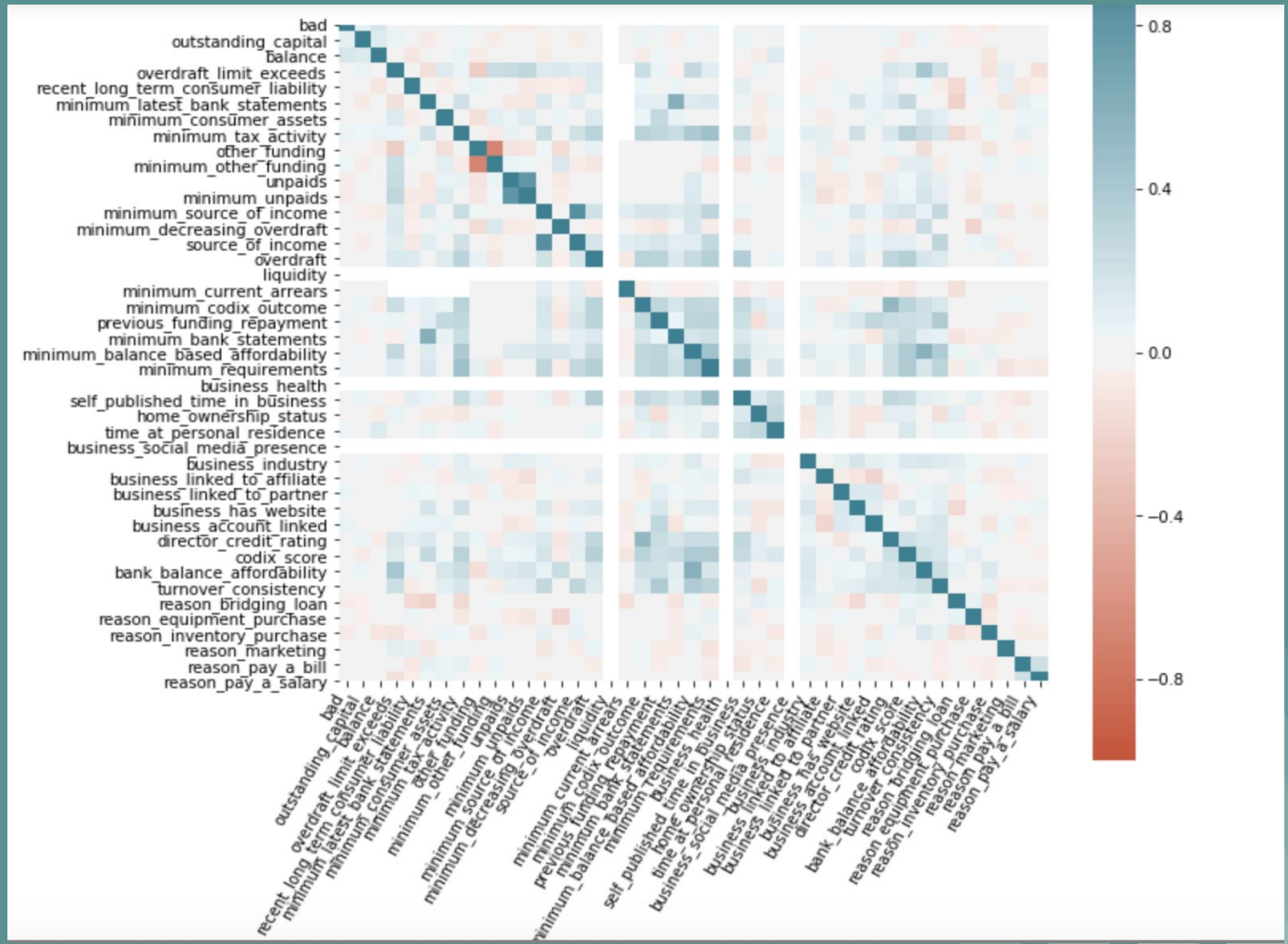
Data Exploratory

- The dataset consists of 1346 rows representing 1346 companies who are Lulalend's customers and 49 columns.
- The target variable is 'bad' which is binary variable containing the history payment of these companies.
- This data set has continuous and categorical data along with the missing values.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1346 entries, 0 to 1345
Data columns (total 49 columns):
AssessmentRunId                          1346 non-null int64
province                                 1346 non-null object
AdvanceId                                1346 non-null int64
bad                                      1346 non-null int64
outstanding_capital                      1346 non-null float64
balance                                  1346 non-null float64
overdraft_limit_exceeds                  391 non-null float64
recent_long_term_consumer_liability      764 non-null float64
minimum_latest_bank_statements           820 non-null float64
minimum_consumer_assets                  820 non-null float64
minimum_tax_activity                     820 non-null float64
other_funding                            1247 non-null float64
minimum_other_funding                    1248 non-null float64
unpaids                                  1247 non-null float64
minimum_unpaids                          1248 non-null float64
minimum_source_of_income                 1248 non-null float64
minimum_decreasing_overdraft             1248 non-null float64
source_of_income                         1247 non-null float64
overdraft                                1247 non-null float64
liquidity                                1247 non-null float64
minimum_current_arrears                  1248 non-null float64
minimum_codix_outcome                    1191 non-null float64
previous_funding_repayment               1286 non-null float64
minimum_bank_statements                  1346 non-null int64
minimum_balance_based_affordability      1346 non-null int64
minimum_requirements                     1346 non-null int64
business_health                          98 non-null float64
self_published_time_in_business         1345 non-null float64
home_ownership_status                    1345 non-null float64
time_at_personal_residence               1345 non-null float64
business_social_media_presence           1345 non-null float64
business_industry                        1345 non-null float64
business_linked_to_affiliate            1345 non-null float64
business_linked_to_partner               1345 non-null float64
```

Variables Selection

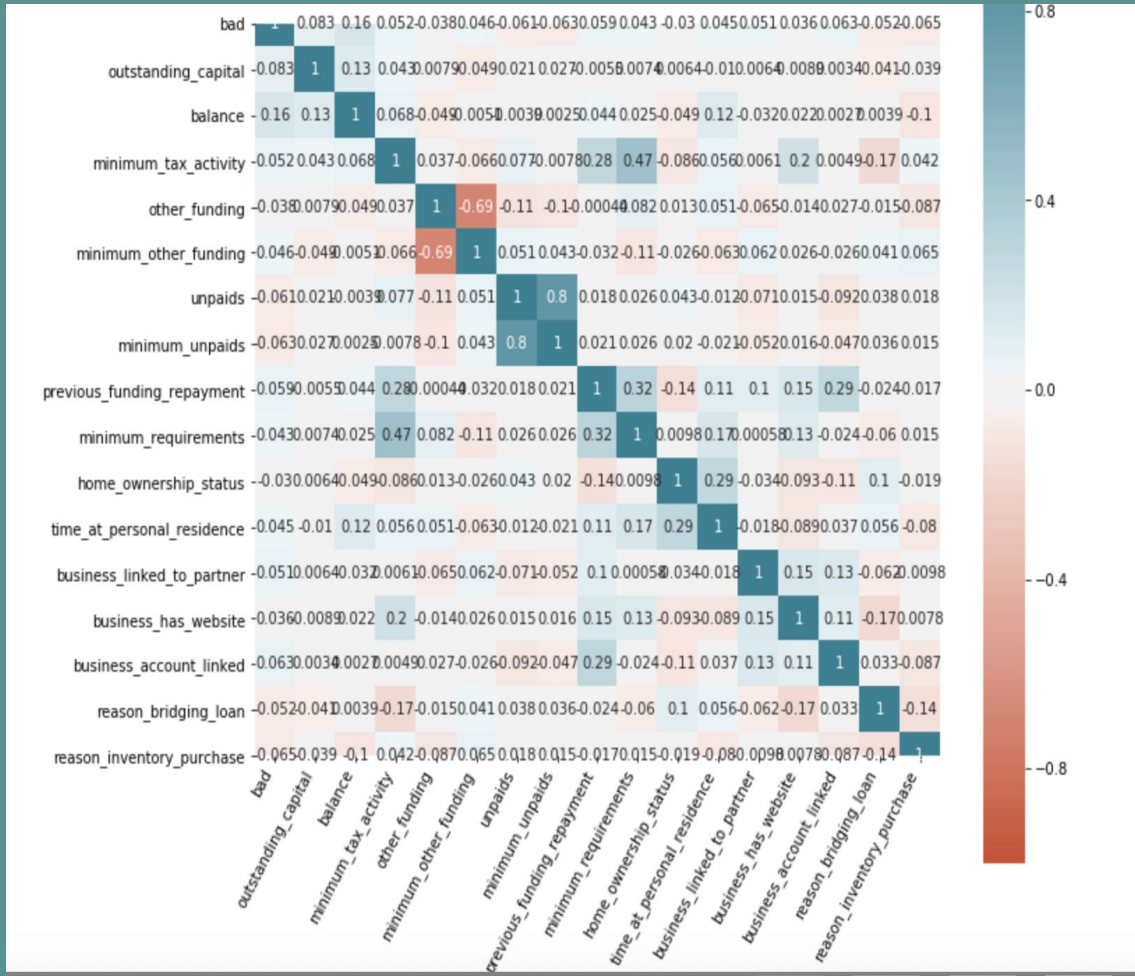
- Not all variables (features) is going to have an impact on the target 'bad'.
- There are 16 variables that have the absolute value of correlation with the target ‘bad’ more than 0.03.



Variables Selection

- We will look at the correlations between these 16 variables to find out the one are highly correlated.

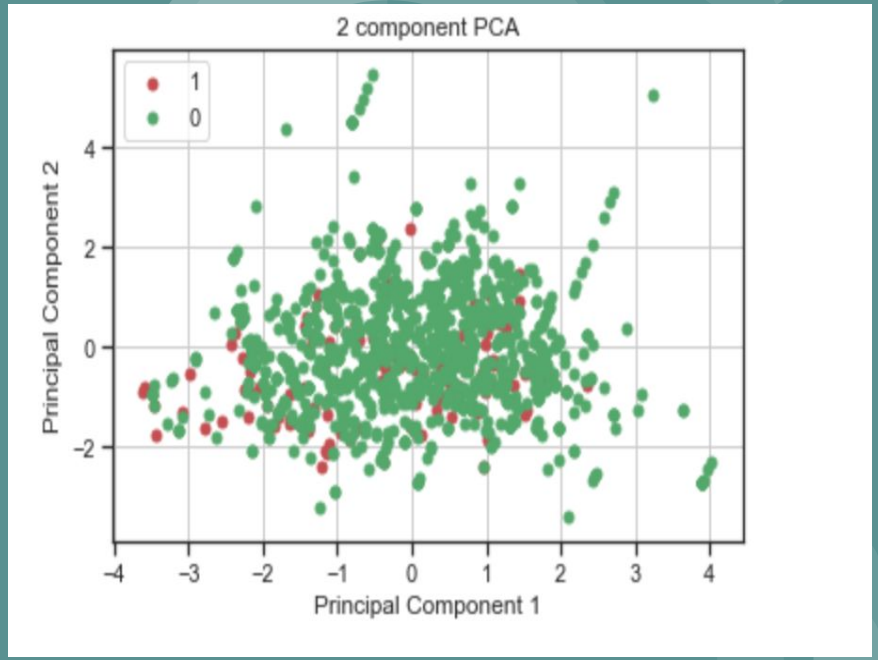
- Three pairs of variables having high correlation:
  - Minimum other funding & other funding
  - Minimum unpaids & Unpaids
  - Minimum requirements & Minimum tax activities



Dimension Reduction PCA

- We use PCA (Principal Component Analysis) to rotate and project data along the direction of increasing variance.

- The first principal component explains 12.7% of the variances, the second principal component explains 11% of the variances.



Analyze by Pivoting Features

- Group minimum_tax_activity = 1 is more likely to miss payments.
- Group minimum_other_funding = 1 is more likely to miss payments.
- Group minimum_unpaids = 0 is more likely to miss payments.

	minimum_tax_activity	bad
1	1.0	0.120773
0	0.0	0.088670

	minimum_other_funding	bad
1	1.0	0.117816
0	0.0	0.078431

	minimum_unpaids	bad
0	0.0	0.161677
1	1.0	0.103608

Analyze by Pivoting Features

- The group home_ownership_status = 1 has the lowest rate of missing payments. It means the directors who own 100% their house are less likely to miss payments.
- The longer the director live in his home, the higher the rate of missing payments.

	home_ownership_status	bad
1	0.25	0.131034
2	0.50	0.121053
0	0.00	0.114600
3	1.00	0.076503

	time_at_personal_residence	bad
4	1.00	0.128866
3	0.75	0.126437
2	0.50	0.106061
0	0.00	0.095238
1	0.25	0.083333

Analyze by Pivoting Features

- The group minimum_tax_activity = 1 is more likely to miss payments.
- The group business_has_website = 1 tend to miss payments more but the difference is very slight.
- The group business_account_linked = 1 tend to miss payments more.

	business_linked_to_partner	bad
1	1.0	0.177215
0	0.0	0.109005

	business_has_website	bad
1	1.0	0.123306
0	0.0	0.100494

	business_account_linked	bad
1	1.0	0.131653
0	0.0	0.091918

Analyze by Pivoting Features

- The group getting the loan for bridging_loan purpose is less likely to miss the payments.
- The group getting the loan for inventory purpose is less likely to miss the payments.

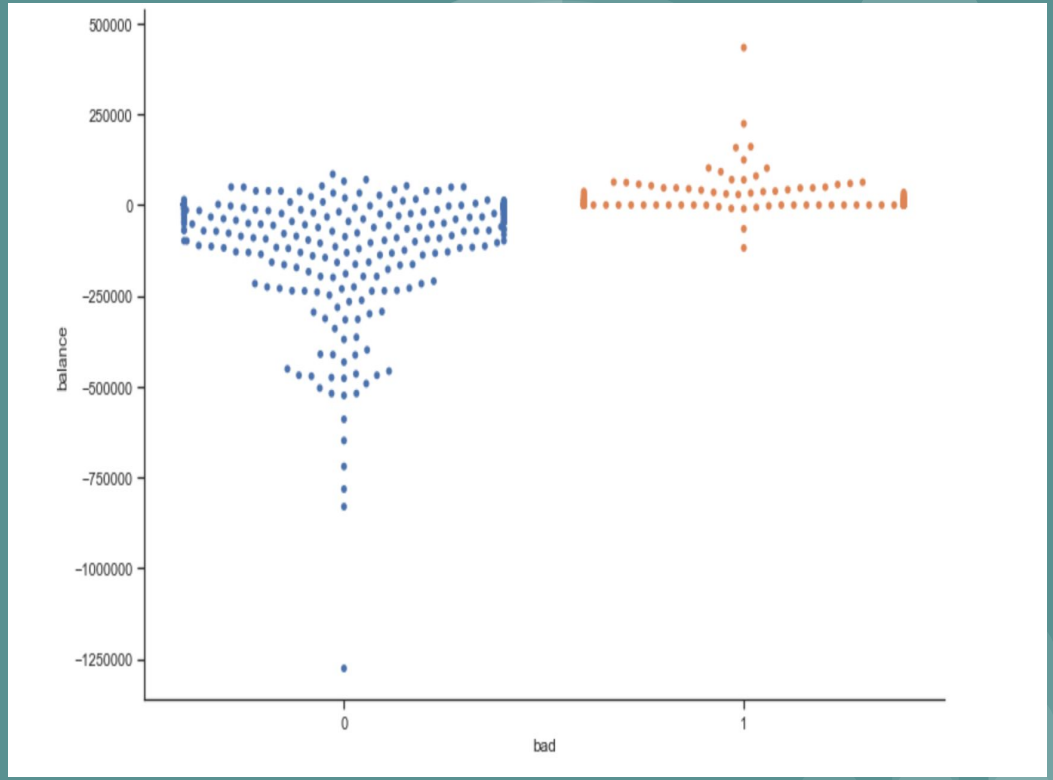
	reason_bridging_loan	bad
0	0	0.126136
1	1	0.087983

	reason_inventory_purchase	bad
0	0	0.124844
1	1	0.095413

Data Visualization

Observations in the group of missing data are more likely to have positive 'balance'.

Which makes sense here since company with arrears tend to miss payments more.

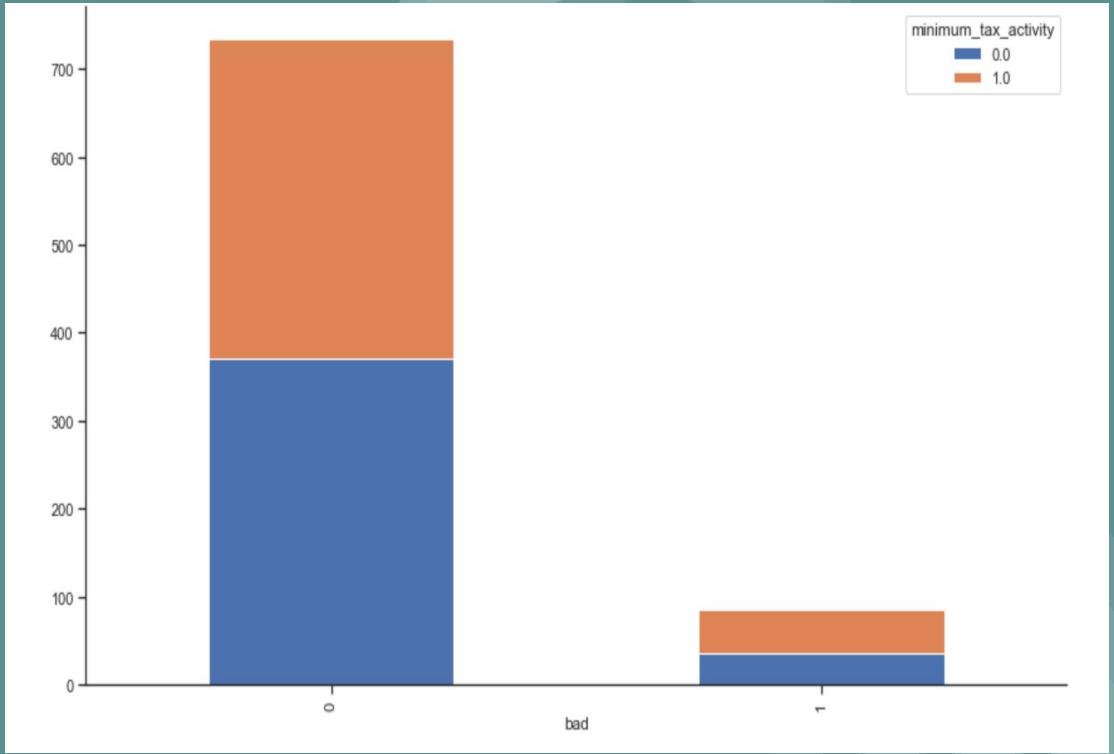


Data Visualization

In the group of not missing payments, 49.6% submitted tax return('minimum_tax_activity' = 1).

While in the group of missing payments, 58% submitted tax return.

The group of missing data is more likely to submit tax return

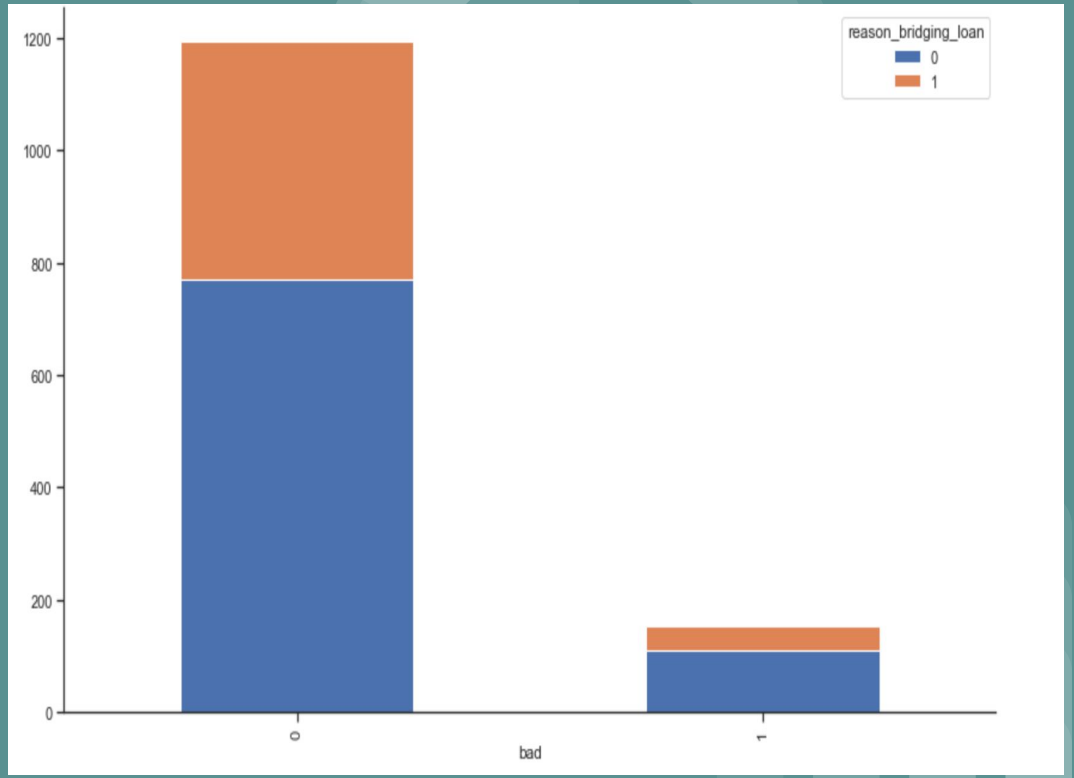


Data Visualization

In the group of not missing payments, 35.59% get the loan for bridging reason.

In the group of missing payments, 26.97% get the loan for bridging reason.

The group of not missing payments tend to get the loan for bridging purpose more.

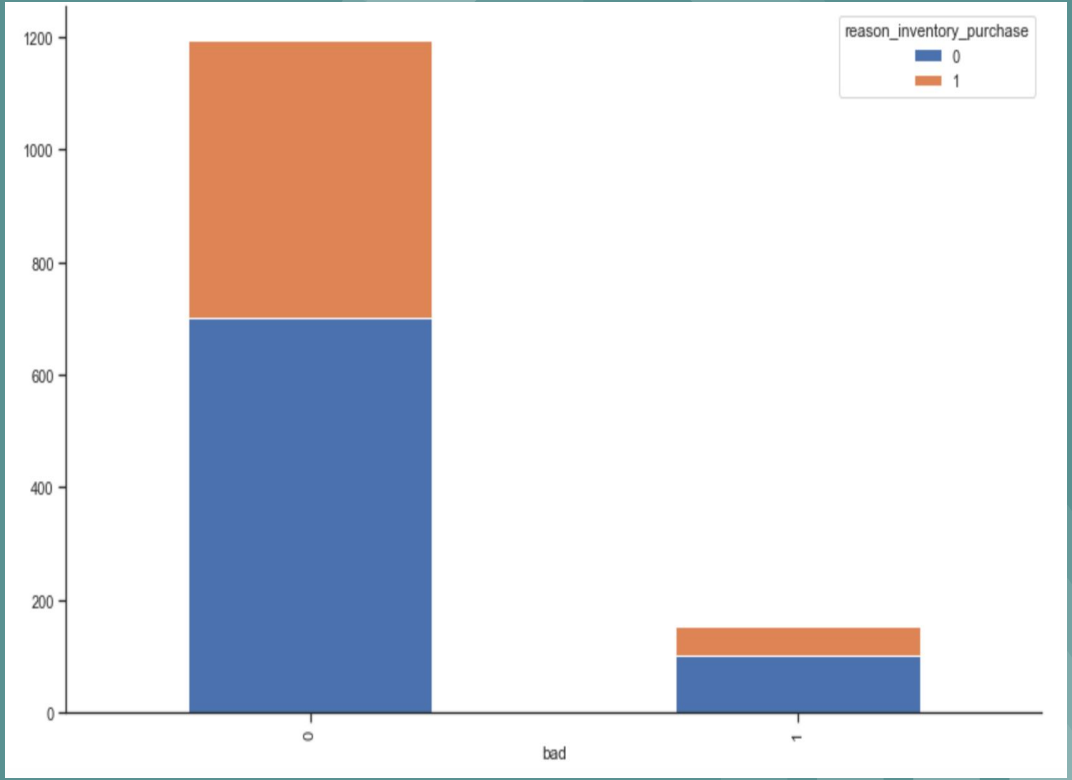


Data Visualization

In the group of not missing payments, 41.29% get the loan for inventory reason.

In the group of missing payments, 34.21% get the loan for inventory purchase reason.

The group of not missing payments tend to get the loan for inventory purchase purpose more.



Machine Learning

- Since the dataset is unbalanced (11.5% of the data labeled 1 and 88.5% labeled 0), we are going to apply the Machine Learning Algorithms on the original training set, oversampling training set and undersampling training set.

- Our problem is a classification problem. We are also performing a category of machine learning which is called supervised learning as we are training our model with a labeled dataset.

- We are using the following Machine Learning Models: K-Nearest Neighborhood, Logistic Regression, Support Vector Machine, Stochastic Gradient Descent, Random Forest, Decision Tree, Naive Bayes.

- The metrics to compare these models are: Accuracy, Precision, Recall, F1-score and area under ROC (receiver operating characteristic) curve.

K-Nearest Neighbor

	Accuracy	Precision	Recall	F1-Score	AUC
Original Data	0.93	0.93	0.93	0.91	0.7503
Oversampling	0.84	0.86	0.84	0.85	0.6912
Undersampling	0.78	0.86	0.78	0.81	0.7234

K-Nearest Neighbor model works the best on original data with high accuracy, precision, recall, f1-score and AUC.

Logistic Regression

	Accuracy	Precision	Recall	F1-Score	AUC
Original Data	0.91	0.91	0.91	0.88	0.7325
Oversampling	0.92	0.91	0.92	0.91	0.7402
Undersampling	0.92	0.91	0.92	0.91	0.7601

Logistic Regression model works the best on undersampling data with high accuracy, precision, recall, f1-score and AUC.

Support Vector Machine

	Accuracy	Precision	Recall	F1-Score	AUC
Original Data	0.93	0.93	0.93	0.91	0.7698
Oversampling	0.87	0.88	0.87	0.87	0.7448
Undersampling	0.92	0.91	0.92	0.91	0.7422

Stochastic Gradient Descent

	Accuracy	Precision	Recall	F1-Score	AUC
Original Data	0.93	0.93	0.93	0.92	7.009
Oversampling	0.86	0.87	0.86	0.87	0.6997
Undersampling	0.55	0.85	0.55	0.63	0.6137

Stochastic Gradient Descent model works the best on original data with high accuracy, precision, recall, f1-score and AUC.

Decision Tree

	Accuracy	Precision	Recall	F1-Score	AUC
Original Data	0.94	0.94	0.94	0.92	0.7277
Oversampling	0.79	0.87	0.79	0.82	0.7567
Undersampling	0.74	0.85	0.74	0.78	0.6804

Decision Tree model works the best on original data with high accuracy, precision, recall, f1-score and AUC.

Random Forest

	Accuracy	Precision	Recall	F1-Score	AUC
Original Data	0.91	0.9	0.91	0.9	0.7403
Oversampling	0.87	0.87	0.87	0.87	0.7295
Undersampling	0.76	0.84	0.76	0.8	0.653

Random Forest model works the best on original data with high accuracy, precision, recall, f1-score and AUC.

Naive Bayes

	Accuracy	Precision	Recall	F1-Score	AUC
Original Data	0.88	0.79	0.88	0.83	0.7414
Oversampling	0.26	0.91	0.26	0.29	0.737
Undersampling	0.91	0.89	0.91	0.89	0.7257

Naive Bayes model works the best on original data with high accuracy, precision, recall, f1-score and AUC.

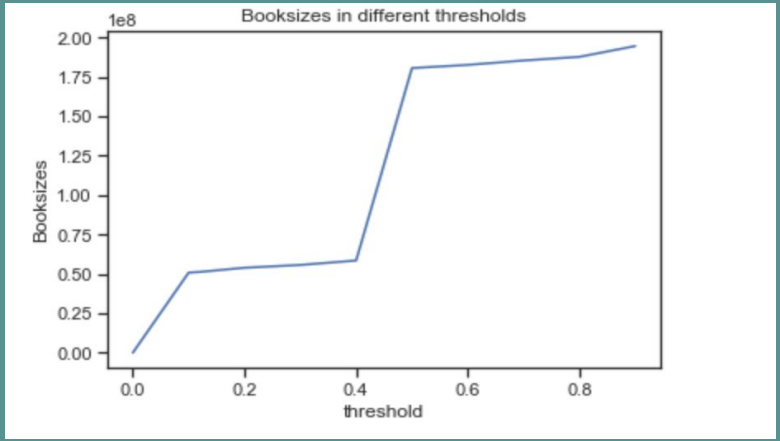
Result

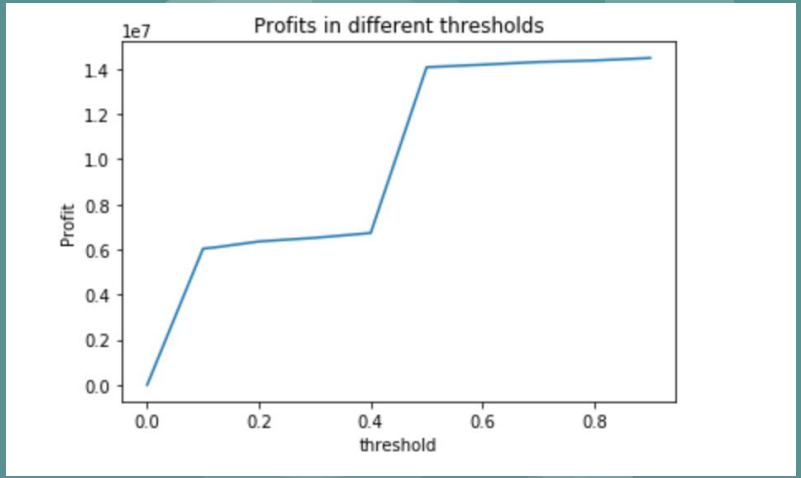
	Accuracy	Precision	Recall	F1-Score	AUC
KNN Original Data	0.93	0.93	0.93	0.91	0.7503
Logistic Regression Undersampling	0.92	0.91	0.92	0.91	0.7601
SVC Undersampling	0.92	0.91	0.92	0.91	0.7422
SGD Original Data	0.93	0.93	0.93	0.92	7.009
Decision Tree Original Data	0.94	0.94	0.94	0.92	0.7277
Random Forest Original Data	0.91	0.9	0.91	0.9	0.7403
Naive Bayes Undersampling	0.91	0.89	0.91	0.89	0.7257

Disbursement & Threshold

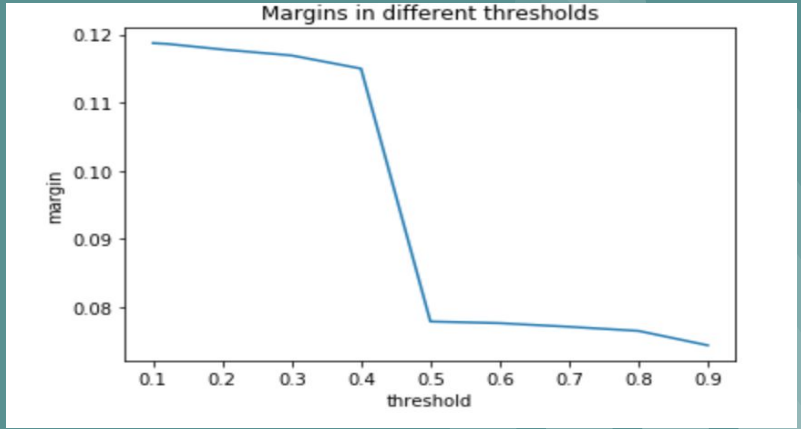
- The Logistic Regression Model provides the probability that a company is going to miss payments. By default, the cut off is 0.5. If a company has more than 50% to miss payments, it is labeled 1 (target feature ‘bad’ is predicted to be 1) .
- The disbursement amount requested by each company is given.
- Is 0.5 the best cut off for the lender?
- With interest rate = 12%, we are going to calculate the booksize, profit and margin at different threshold to figure out which threshold works the best for lender.

Disbursement & Threshold





- When the threshold is 0.1, the margin is 0.119 and the profit is 6,039,308.
- When the threshold is 0.4, the margin is 0.115 and the profit is 6,731,253.
- The margin at 0.4 is slightly less than the best one, but the profit is higher. The best cut off is 0.4.



Conclusion

- The lender can use the best model to find the probability of missing payments of a future customer.
- Using the probability of missing payments, disbursement amount and other factors, lender can make decision whether to fund or not.
- They also can find the best threshold at different interest rate.