Explainable Machine Learning: Credit Scoring for Home Equity Borrowers

Machine Learning Nanodegree Capstone Project

Shivraj Bheenick
August 25, 2018

I. Definition

Project Overview

Credit scores are used to assess consumer credit risk in terms of the likelihood of repaying a borrowed amount. Lenders use the credit scores to assess how much credit should be granted to a borrower, if at all, and what are the terms and interest rates for any credit extended. As well as the accuracy of credit scoring methodologies, there is an equally pressing need for regulators, consumers and credit analysts to interpret the models used in credit scoring. While regulators are keen to supervise scoring practices to ensure fairness, consumers are entitled to know the basis for their individual credit score and how this can be improved. Finally, interpretability of the models will also help credit analysts understand their datasets and the models' predictions, detect and correct for biases, and ultimately create better models.

FICO has provided an anonymised dataset of roughly 10 thousand HELOC applications made by real homeowners. The dataset contains 5,000 "Good" records where borrowers repaid as negotiated over the first 12-36 months of their loan. There are 5,459 "Bad "records where the borrower made a seriously late payment in that same period. Thus, the target variable (Good/Bad loan) is equally represented in the dataset, which has 23 features relating to different aspects of creditworthiness such as length of credit, payment history and amount of debt. A data dictionary file has also been provided to explain each feature and its relationship to the target variable, which is captured by the monotonicity constraint.

Problem Statement

As part of the FICO Explainable Machine Learning Challenge, machine learning techniques will be used to assess the credit risk of home equity borrowers who are seeking a Home Equity Line of Credit (HELOC)[1] in the range of $5,000 - $150,000. Each borrower will be assigned a binary RiskPerformance classification. A "Bad" RiskPerformance indicates that a consumer was 90 days past due or worse at least once over a period of 24 months from when the credit account was opened, while a "Good" value implies the borrower has made his payments without ever being more than 90 days overdue. This classification will help determine whether the homeowner qualifies for a line of credit and, if so, how much credit should be extended.

[1] A HELOC is a line of credit typically offered by a bank as a percentage of home equity (the difference between the current market value of a home and its purchase price)

The two key objectives of this project are to maximise the prediction performance of the model while retaining the interpretability of the model. To meet the latter requirement, two types of explanations will be generated. Firstly, a global explanation will provide transparency over the overall prediction mechanism of the model, including which features are the most important in determining the credit score. A local explanation will also be generated for each customer prediction to justify individual credit scores and identify how the customer can improve his credit score to have a better chance of getting a loan in the future.

The project will be structured as follows:

a. **Exploratory Data Analysis:** The individual features will be analysed via visualisations to discern any correlations and patterns with the target variable. Data imputation will be performed to fix any gaps. New features will potentially be derived off the existing features if they are deemed to be relevant for predicting RiskPerformance. The importance of the features will be assessed via PCA.
b. **Model Explainability Approach:** For each of the chosen models, a suitable framework will be put together to generate local and global explanations. This will be a mixture of existing sklearn functions (e.g feature importance) as well as more sophisticated approaches (e.g LIME).
c. **Model Fitting and Optimisation:** Each of the 3 models (logistic regression, decision tree classifier, gradient boosting classifer) will be fitted to the dataset using cross-fold validation and the model hyperparameters will be tuned to maximise performance.
d. **Evaluation of Optimised Models:** The results, including the explanations, from the optimised models will be analysed in the context of the credit scoring exercise. The recommended model will be chosen based on its prediction performance, the simplicity and transparency of its explanations and the plausibility of its interpretations.

Metrics

In the context of the HELOC dataset, both false negatives and false positives lead to unfavourable outcomes. False negative results in the refusal of credit to a good customer, with the subsequent loss of interest income. This will be monitored via the *sensitivity* measure.

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

False positive results in incorrectly classifying a bad loan as good and thereby granting credit to a bad customer on favourable terms. This can result in loss of the loan principal as well as any additional costs arising from the loan default. This will be monitored via the *specificity* measure.

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

Both sensitivity and specificity will be combined in a single score (G-mean). The larger the G-mean, the superior is the model. This is the metric that will be optimised for the various machine learning algorithms used.

$$G = \sqrt{sensitivity\ \times specificity}$$

To measure the interpretability of the models, global and local explanations will be generated. The 2 models investigated comprise of decision trees – the number of features used will be restricted (less than 10) together with the maximum depth of these models to avoid overly complex models. The feature importance will be visualised and interpreted to explain the prediction models.

II. Analysis

Data Exploration

FICO uses 5 broad categories of information to determine credit score of borrowers. The 23 features provided in the input dataset have been subjectively mapped to these categories below to develop a better intuition of the input features.

- **Payment History**: Considers payment history from different types of accounts, public record and collection items and details on late or missed payments.
    - ***NumSatisfactoryTrades***: Number of credit agreements with on-time payments
    - ***NumTrades60Ever2DerogPubRec***: Number of trade lines on a credit bureau report that record a payment received 60 days past its due date. This feature also checks all Public Records available for the consumer and adds to this count any items considered "Derogatory".
    - ***NumTrades90Ever2DerogPubRec***: Number of trade lines on a credit bureau report that record a payment received 90 days past its due date. This feature also checks all Public Records available for the consumer and adds to this count any items considered "Derogatory".
    - ***PercentTradesNeverDelq***: Percent Trades Never Delinquent
    - ***MSinceMostRecentDelq***: Months Since Most Recent Delinquency
    - ***MaxDelq2PublicRecLast12M***: Categorical variable denoting how severe was delinquency in last 12 months, if at all and known
    - ***MaxDelqEver***: Categorical variable denoting how severe was delinquency in the past, if at all and known

- **Amount of Debt**: Considers total amount owed across all accounts, amounts owed on specific types (revolving, credit) of accounts, number of accounts with balance, credit utilization ratio on revolving accounts, remaining amount owed on instalment loans
    - ***NetFractionRevolvingBurden***: Revolving balance divided by credit limit
    - ***NetFractionInstallBurden***: Instalment balance divided by original loan amount
    - *NumBank2NatlTradesWHighUtilization*: Number of credit cards carrying a balance at 75% of its limit or greater
    - ***PercentTradesWBalance***: Percent Trades with Balance

- **Length of Credit History**: Considers age of oldest account, average account age and age of specific types of accounts (credit card, auto loans, etc)
    - ***MSinceOldestTradeOpen***: Months Since Oldest Trade Open
    - ***AverageMInFile***: Average history length of trades in months

- **New Credit**: Considers the number of new accounts, how long since new account opened, number of recent requests for credit and rate shopping for a single loan
    - ***MSinceMostRecentTradeOpen***: Months Since Most Recent Trade Open
    - ***NumTradesOpeninLast12M***: Number of Trades Open in Last 12 Months
    - ***MSinceMostRecentInqexcl7days***: Months Since Most Recent Inquiry excluding 7days

    - ***NumInqLast6M***: Number of Inquiry Last 6 Months
    - ***NumInqLast6Mexcl7days***: Number of inquiries Last 6 Months excluding 7days (to account for price comparison shopping)

- **Credit Mix**: Considers types of credit accounts (credit cards, retail accounts, instalment loans and mortgage loans)
    - ***PercentInstallTrades***: Percent Instalment Trades
    - ***NumRevolvingTradesWBalance***: Number Revolving Trades with Balance
    - ***NumInstallTradesWBalance***: Number Instalment Trades with Balance

- **Other Category**: Remaining features which could not be mapped to one of the above FICO categories
    - ***ExternalRiskEstimate***: Consolidated risk estimation from other credit bureaus
    - ***NumTotalTrades***: Number of Total Trades (total number of credit accounts)

Of the 23 features, 21 are numerical with the remaining 2 (MaxDelq2PublicRecLast12M**,** MaxDelqEver) being categorical. A statistical description of the numerical features is provided in Table 1.

Feature	mean	std	min	25%	50%	75%	max
ExternalRiskEstimate	67.43	21.12	-9	63	71	79	94
MSinceOldestTradeOpen	184.21	109.68	-9	118	178	250	803
MSinceMostRecentTradeOpen	8.54	13.30	-9	3	5	11	383
AverageMInFile	73.84	38.78	-9	52	74	95	383
NumSatisfactoryTrades	19.43	13.00	-9	12	19	27	79
NumTrades60Ever2DerogPubRec	0.04	2.51	-9	0	0	1	19
NumTrades90Ever2DerogPubRec	- 0.14	2.37	-9	0	0	0	19
PercentTradesNeverDelq	86.66	26.00	-9	87	96	100	100
MSinceMostRecentDelq	6.76	20.50	-9	-7	-7	14	83
NumTotalTrades	20.86	14.58	-9	12	20	29	104
NumTradesOpeninLast12M	1.25	3.07	-9	0	1	3	19
PercentInstallTrades	32.17	20.13	-9	20	31	44	100
MSinceMostRecentInqexcl7days	- 0.33	6.07	-9	-7	0	1	24
NumInqLast6M	0.87	3.18	-9	0	1	2	66
NumInqLast6Mexcl7days	0.81	3.14	-9	0	1	2	66
NetFractionRevolvingBurden	31.63	30.06	-9	5	25	54	232
NetFractionInstallBurden	39.16	42.10	-9	-8	47	79	471
NumRevolvingTradesWBalance	3.19	4.41	-9	2	3	5	32
NumInstallTradesWBalance	0.98	4.06	-9	1	2	3	23
NumBank2NatlTradesWHighUtilization	0.02	3.36	-9	0	0	1	18
PercentTradesWBalance	62.08	27.71	-9	47	67	82	100

Table 1: Descriptive statistics of numerical features

The 2 categorical features have their values explained in Table 2.

Feature	Value	Meaning
MaxDelq2PublicRecLast12M	0	derogatory comment
	1	120+ days delinquent
	2	90 days delinquent
	3	60 days delinquent
	4	30 days delinquent
	5, 6	unknown delinquency
	7	current and never delinquent
	8, 9	all other
MaxDelqEver	1	No such value
	2	derogatory comment
	3	120+ days delinquent
	4	90 days delinquent
	5	60 days delinquent
	6	30 days delinquent
	7	unknown delinquency
	8	current and never delinquent
	9	all other

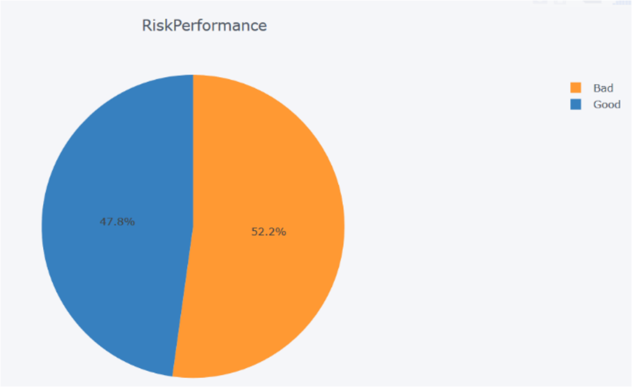
Table 2: Interpretation of categorical features

There are also 3 special values in the dataset:

- -9 (No Bureau Record or No Investigation): There are 588 records where all features have this value. The confounding of no bureau report investigated (most likely a VIP applicant) and no bureau report found (a negative trait for extending credit) is most likely responsible for such loans being classified as Good and Bad. Given there is no usable information for these records, they will be excluded from the dataset during the data pre-processing stage.
- -8 (No Usable/Valid Trades or Inquiries): The NetFractionInstallBurden feature displays a large gap in terms of usable/valid trades.
- -7 (Condition not Met e.g. No Inquiries, No Delinquencies): The MSinceMostRecentDelq feature has 4,600 entries with this special value, reflecting no previous delinquency. This might be a very useful piece of information in determining RiskPerformance.

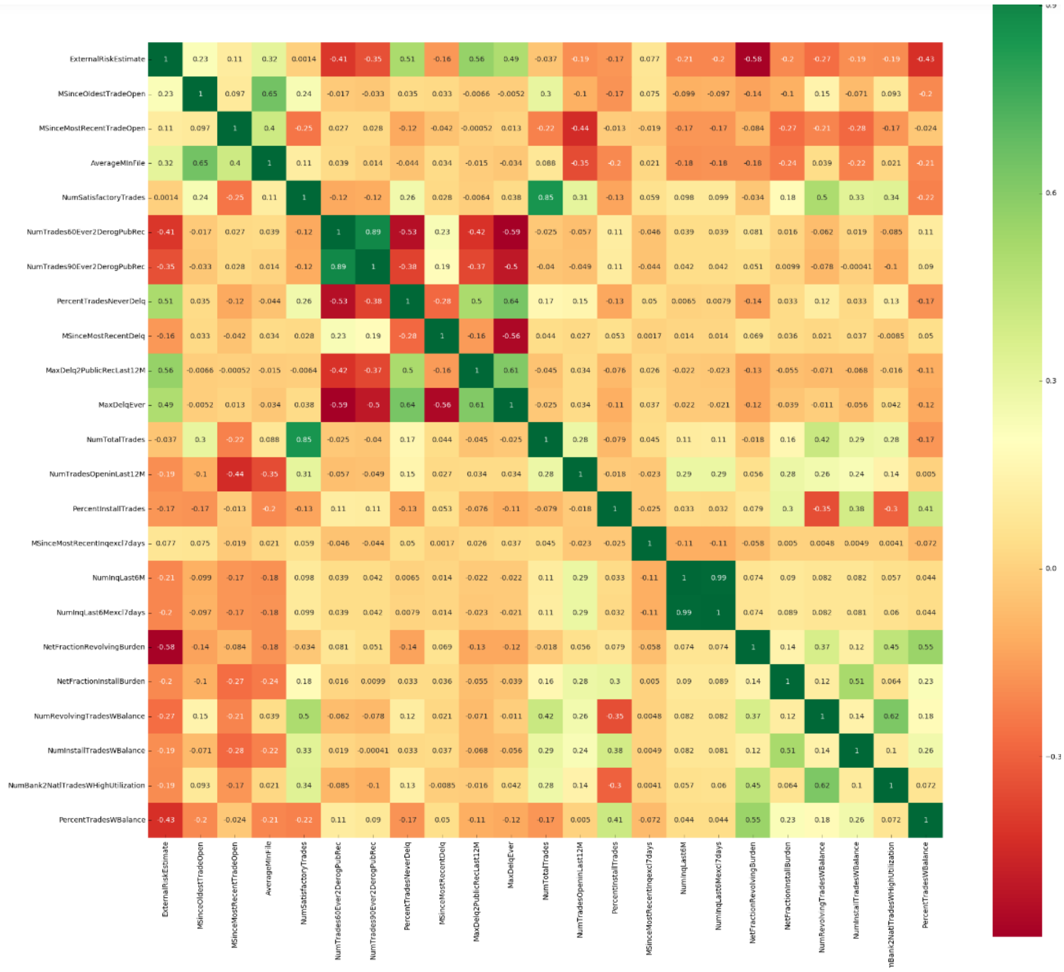
Exploratory Visualization

Chart 1: RiskPerformance distribution



Both the target variable (Good, Bad) are equally represented in the dataset; this balance will be preserved during the random sampling of data for model training.

Chart 2: Correlation of features

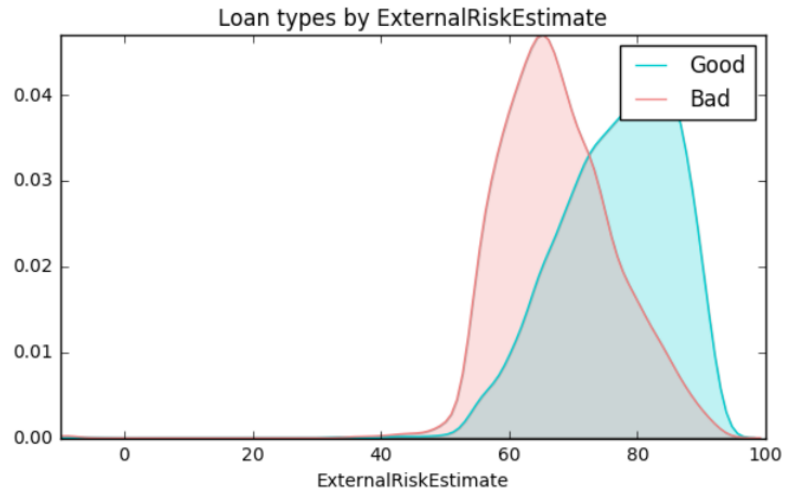


The following pairs of features exhibit strong correlation:

NumInqLast6M and NumInqLast6Mexcl7days: NumInqLast6M will be dropped from the feature set as NumInqLast6Mexcl7days accounts for the borrower doing multiple credit inquiries to get the best loan rate.

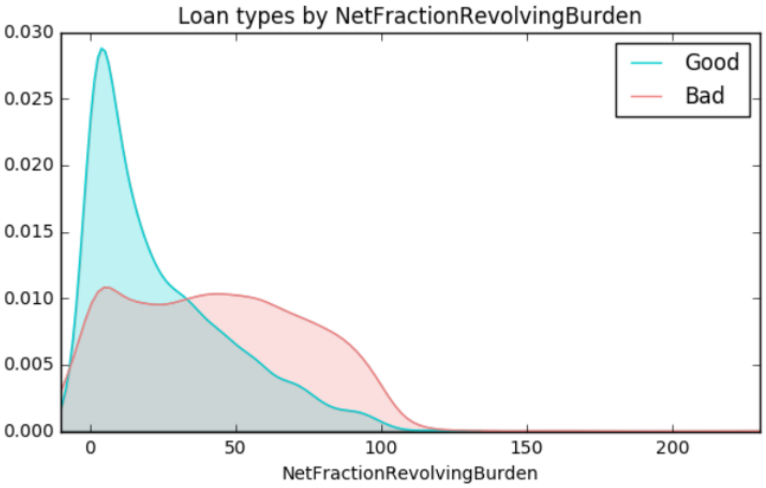
- NumTrades60Ever2DerogPubRec and NumTrades90Ever2DerogPubRec: NumTrades90Ever2DerogPubRec will be dropped from the feature set as it is a subset of NumTrades60Ever2DerogPubRec
- NumSatisfactoryTrades and NumTotalTrades: NumTotalTrades will be dropped after it has been used to derive features that are expressed as a % of number of trades (PercentTradesNeverDelq, PercentInstallTrades, PercentTradesWBalance)

Chart 3: RiskPerformance vs ExternalRiskEstimate



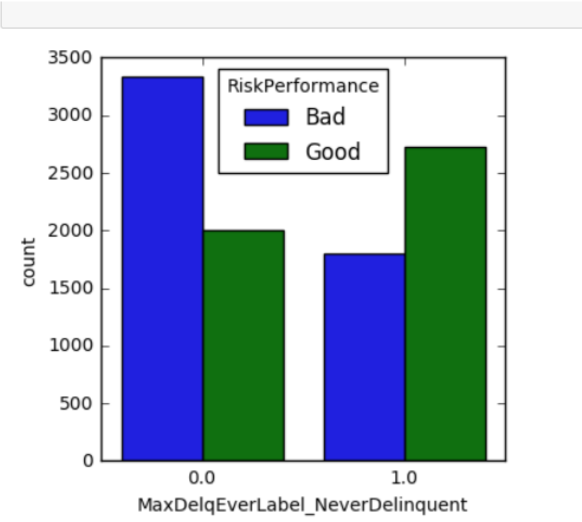
This density plot confirms that RiskPerformance is predominantly good as the ExternalRiskEstimate increases, with the crossover at a score of around 70. Of all the features in the dataset, ExternalRiskEstimate is the least interpretable as no insight is provided into how it is computed.

Chart 4: RiskPerformance vs NetFractionRevolvingBurden



This density plot shows that borrowers who use more than 40% of their credit limit are more likely to make a seriously late payment than those who have a lower relative debt levels.

Chart 5: RiskPerformance vs Past Delinquency



This bar chart shows that borrowers who had never been delinquent (label 1) in the past are more likely repay on time, while those with past delinquency in the last 12 months or before (label 0) are more likely to be in arrears again.

Algorithms and Techniques

Models

The categorisation of home equity loans into one of two types of RiskPerformance is a binary classification problem for which the following 3 supervised learning techniques will be used:

- **Logistic Regression (Benchmark model):** This uses the function below to produce output a set of probabilities (between 0 and 1) for the target outcomes. One of the main advantages of logistic regression for the credit scoring classification problem is that it allows the threshold probablity to be adjusted for a loan to be classified as Good/Bad.

$$P(y_i = 1) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p}))}$$

- **Decision Tree Classifier:** Unlike logistic regression, decision trees also work for cases where the relationship between features and the outcome is non-linear or where the features interact with each other. The decision tree repeatedly splits the data into smaller subsets using cut-off values based on the features to minimise the Gini Index. The last splits result in the data subset being assigned to one of the target outcomes.
- **Gradient Boosting Classifier:** This is an ensemble model that will make use of many shallow decision trees and will sequentially focus on the reduction of the mean square error to increase the prediction performance. After the initial tree is grown, each tree in the series is

fitted with the purpose of reducing the error, hence we would expect a better score than an individual decision tree classifier.

An additional model (logistic model tree) was put forward in the project proposal; this has been subsequently dropped as there was no mature implementation of this model which was available, and it was too tedious to build the model from scratch within the capstone timelines.

Global Interpretation

The ***feature importance*** metric of sklearn was used to determine which features influenced the prediction the most. For both the tree models used, the depth of a feature used as a decision node in a tree is used to assess the relative importance of that feature. The higher the feature in the tree, the more input samples are likely to be predicted using that feature. Thus, the relative importance of a feature is determined by the expected fraction of the samples they contribute to. For the gradient boosting classifier (and any other ensemble of trees), the overall feature importance is determined by simply averaging the feature importance of each individual tree.

Local Interpretation

Local interpretable model-agnostic explanations (LIME) will be generated to explain which features contributed to the prediction of individual loans. To generate an explanation, LIME will perturb the features of the loan instance to be explained, use the black-box machine learning model to predict the outcomes for the perturbed instances and fit a linear model to these predictions. The feature importance is based on the linear model fitted and is only faithful locally to the original loan instance that needs to be explained.

Benchmark

Given the dual goal of accuracy and interpretability, the logistic regression model will be used as benchmark as the odds ratio can be used to interpret its output. Logistic regression is widely used for predicting credit scores and a minimum Gmean score of 65% will be expected for the classification of the RiskPerformance for the HELOC dataset.

III. Methodology

Data Preprocessing

The 588 loan records where all features had a value of -9 were imputed from the dataset as they offered no value for training the model. While the features had different order of magnitudes, no feature scaling was carried out to preserve the real values which would facilitate the interpretation of the results. The target variable RiskPerformance was mapped to 0 for Good and 1 for Bad loans.

Feature Engineering

The 2 categorical features (MaxDelq2PublicRecLast12M*,* MaxDelqEver) had their values mapped to one of the following categories and subsequently one-hot encoded to generate distinct features:

- Derogatory
- Delinquent (30, 60, 90, 120+ days)
- Never delinquent (current)
- Invalid (none of the above categories)

3 additional features were also derived using the raw features in the dataset:

- **NumTradesNeverDelq**: This feature relates to the payment history of the borrower and is calculated from the PercentTradesNeverDelq and NumTotalTrades.
- **NumInstallTrades**: This feature relates to the credit mix of the borrower and is calculated from PercentInstallTrades and NumTotalTrades.
- **NumTradesWBalance**: This feature relates to the amount of debt of the borrower and is calculated from PercentTradesWBalance and NumTotalTrades.

Final Feature Selection

With the feature engineering above, the total number of features grew from 23 to 35. 3 feature selection techniques were applied to identify the key features for training the supervised models:

a. *Principal Component Analysis:* The first 2 principal components explain 85% of the variance in the dataset (Figure 1).

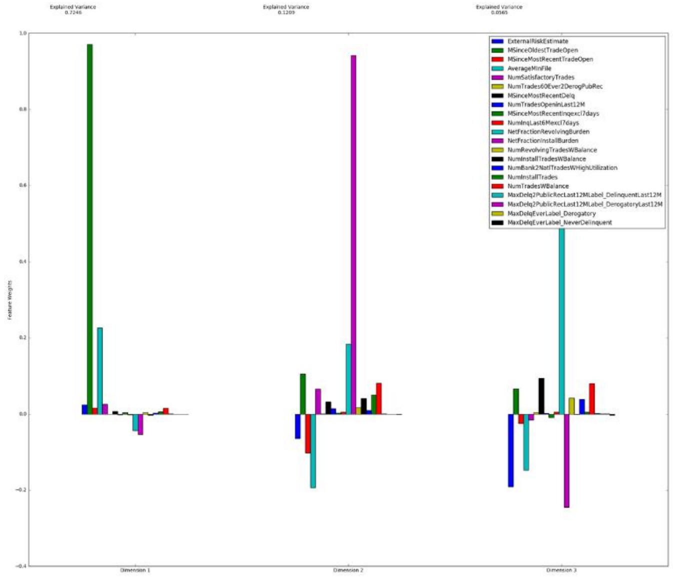


Figure 1: PCA of HELOC feature set

The first principal component (73%variance) captures the length of the credit history of the borrower, with MSinceOldestTradeOpen and AverageMInFile being the 2 most prominent features.

The second component (12% variance) looks at the amount of debt (NetFractionInstallBurden, NetFractionRevolvingBurden) of the borrower as well as the length of his credit history (AverageMInFile)

b. *Recursive Feature Selection:* A maximum Gmean score of 72.8% is obtained when 17 features are used. When ExternalRiskEstimate is excluded from the feature set, the maximum Gmean score drops to 71.7% when 19 features are used. Given 13 features produces an almost identical score (71.5%), it is chosen instead as it will facilitate the interpretation of the predictions due to a lower number of dimensions. These 13 features can be grouped into the following 4 categories:
   - Length of credit History: *AverageMInFile*
   - Payment History: *NumSatisfactoryTrades, NumTrades60Ever2DerogPubRec, MaxDelq2PublicRecLast12MLabel_DelinquentLast12M, MaxDelq2PublicRecLast12MLabel_DerogatoryLast12M, MaxDelqEverLabel_Derogatory, MaxDelqEverLabel_NeverDelinquent*
   - New Credit: *NumTradesOpeninLast12M, MSinceMostRecentInqexcl7days, NumInqLast6Mexcl7days*
   - Amount of debt: *NetFractionRevolvingBurden, NumRevolvingTradesWBalance, NumBank2NatlTradesWHighUtilization*

c. *Tree-Based Feature selection:* An almost perfect Gmean score (99.5%) is achieved for a tree depth of 25, both with and without ExternalRiskEstimate. While the large tree dept is prone to overfitting and hinders model interpretability, this was a useful exercise to identify the most features with the best predictive power, which are categorised below:
   - Amount of debt: *NetFractionRevolvingBurden, NumTradesWBalance*
   - Length of credit History: *MSinceOldestTradeOpen, AverageMInFile*
   - Payment History: *NumSatisfactoryTrades, NumTrades60Ever2DerogPubRec, MSinceMostRecentDelq*
   - New Credit: *MSinceMostRecentInqexcl7days*
   - Credit Mix: *NumInstallTrades*

Based on the insight provided by the above analysis, the following 8 features (ExternalRiskEstimate was deliberately excluded) consistently came up as the most important and were hence selected for training the supervised models.

- Length of Credit History: MSinceOldestTradeOpen, AverageMInFile

- Amount of debt: NetFractionRevolvingBurden, NetFractionInstallBurden (No further manipulation will be done for the 3415 records where there is no data on the burden), NumBank2NatlTradesWHighUtilization
- Payment history: NumSatisfactoryTrades, NumTrades60Ever2DerogPubRec
- New Credit: MSinceMostRecentInqexcl7days

Implementation

The cleaned dataset was split 80:20 into a training and testing set using sklearn's train_test_split function. A Gscore class was created in a separate module to calculate the Gmean score. The code was structured in the following 3 parts:

a. **Training and Optimisation**: The *optimise_model* function:
   - Applies sklearn's StratifiedShuffleSplit to the training set to generate cross-validation datasets
   - Optimises the model using sklearn's GridSearchCV and the parameters supplied by the user
   - Identifies the best model parameters and calculates true positive, false positive, true negative, false negative, sensitivity, specificity and Gmean
   - Invokes the *pred_outcome* function to generate an excel file with the training data together with the actual and predicted outcomes. The analysis of this data for the misclassified records was relatively challenging to identify the root cause of the wrong predictions – this was partly because the analysis only focused on the final feature set instead of the full set of features provided in the initial dataset.

b. **Performance Evaluation:** The *model_predict* function:
   - Calculates the mean score using sklearn's cross_val_score function
   - Invokes the *pred_outcome* function to generate an excel file with the testing data together with the actual and predicted outcomes
   - Calculates true positive, false positive, true negative, false negative, sensitivity, specificity and Gmean

c. **Explainable Framework**
   1. Global Interpretation: sklearn's feature_importances_ method is used to identify the most influential features in predicting RiskPerformance. Given the feature importance is always a positive number between 0 and 1, it was hard to determine whether the monotonicity constraint which had been provided was being respected by the model.
   2. Local Interpretation: LIME was used to generate local explanations for individual loans in the testing set. To fully grasp the contribution of a particular feature in a loan prediction, the feature value has to be assessed relative to its statistical distribution within the dataset. This was carried out manually by referring to Table 1.

Logistic regression, Decision Tree Classifier and Gradient Boosting Classifier were trained, optimised, evaluated and interpreted using the above functions.

Refinement

Each of the 3 models used were optimised using sklearn's GridsearchCV function, with the following parameters being tuned to maximise the Gmean score:

- *Logistic Regression*: {'C':np.arange(1e-05, 0.5, 0.01), 'penalty': ['l1','l2']}
- *Decision Tree Classifier*: {'max_depth':np.arange(2, 21, 4), 'min_samples_leaf':np.arange(50, 301, 50), 'min_samples_split': np.arange(100, 501, 100) }
- *Gradient Boosting Classifier*: {'learning_rate':(0.01, 0.21, 0.1),'n_estimators' :np.arange(100, 501, 200),'max_depth':np.arange(2, 10, 3),'min_samples_leaf':np.arange(50, 301, 100)}

For each model optimised, the model parameters resulting in the highest mean cross-validation score were recorded.

- *Logistic Regression*: Best score is 0.697 using {'C': 0.040010000000000004, 'penalty': 'l1'}
- *Decision Tree Classifier*: Best score is 0.694 using {'min_samples_leaf': 150, 'max_depth': 6, 'min_samples_split': 100}
- *Gradient Boosting Classifier*: Best score is 0.712 using {'learning_rate': 0.1, 'min_samples_leaf': 250, 'max_depth': 2, 'n_estimators': 100}

IV. Results

Model Evaluation and Validation

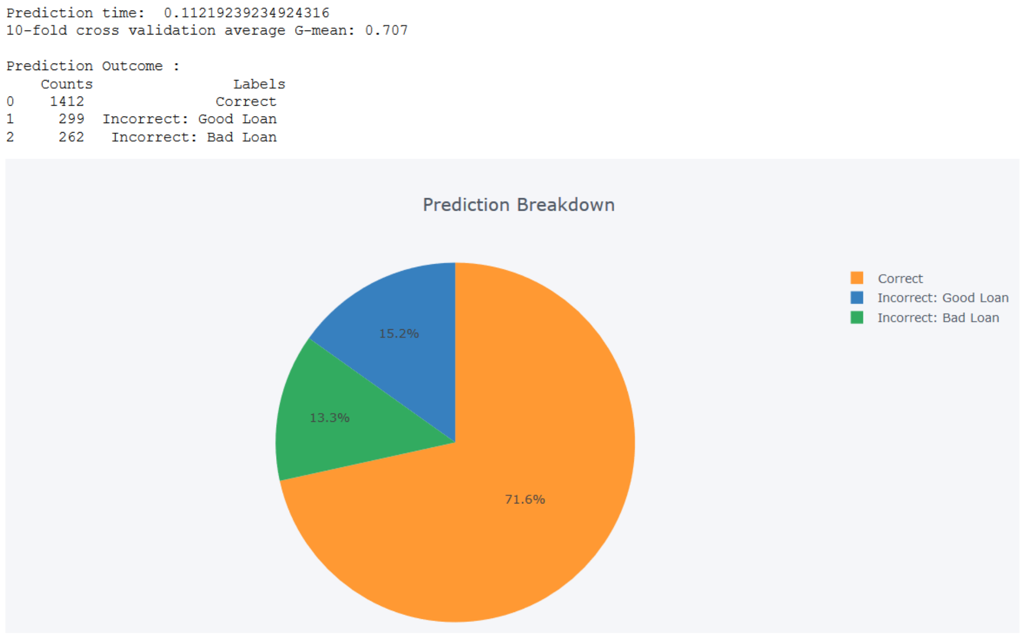
The table below summarises the key metrics for each of the 3 models used on the training as well as the test data. Each model shows similar scores on both datasets, which implies that no overfitting is happening during training.

Model	Run Type	Dataset Size	True Positive	False Positive	True Negative	False Negative	Sensitivity	Specificity	GMean
LogisticRegression	Training	7892	2704	1313	2776	1099	0.711	0.679	0.695
DecisionTreeClassifier	Training	7892	2735	1228	2861	1068	0.719	0.7	0.709
GradientBoostingClassifier	Training	7892	2671	1031	3058	1132	0.702	0.748	0.725
LogisticRegression	Testing	1973	649	346	697	281	0.698	0.668	0.683
DecisionTreeClassifier	Testing	1973	635	315	728	295	0.683	0.698	0.69
GradientBoostingClassifier	Testing	1973	630	262	781	300	0.677	0.749	0.712

Table 4: Performance metrics for model runs

The highest Gmean score (71.2 %,) is obtained by the Gradient Boosting Classifier- this is a respectable score for credit risk classification problems. The optimal model consisted of 100 shallow trees (depth 2) which have at least 250 loans in each leaf node, with the superior predictive power stemming from the ability of new trees to learn from the misclassifications of previous trees.

The 10-fold cross validation average Gmean score for the Gradient Boosting Classifier was 70.7%, which is close to the score obtained above. When used to predict RiskPerformance on the test dataset, this classifier correctly identifies 71.6% of the loans, fails to identify 15.2% of good loans and also fails to detect 13.3% of bad loans.



```
Prediction time:  0.11219239234924316
10-fold cross validation average G-mean: 0.707

Prediction Outcome :
    Counts                Labels
0    1412               Correct
1     299  Incorrect: Good Loan
2     262   Incorrect: Bad Loan
```

Figure 2: Prediction breakdown for optimised Gradient Boosting Classifier

In terms of interpretability, 6 of the 8 selected features had an importance greater than 0.1, with the two features deemed to be the least important being MSinceOldestTradeOpen and NetFractionInstallBurden. The 6 features cover all the key considerations of a creditworthy borrower, notably amount of debt (*NetFractionRevolvingBurden*, *NumBank2NatlTradesWHighUtilization*), length of credit history (*AverageMinFile*), payment history (*NumSatisfactoryTrades*, *NumTrades60Ever2DerogPubRec*) and new credit (*MSinceMOstRecentInqexcl7days*).

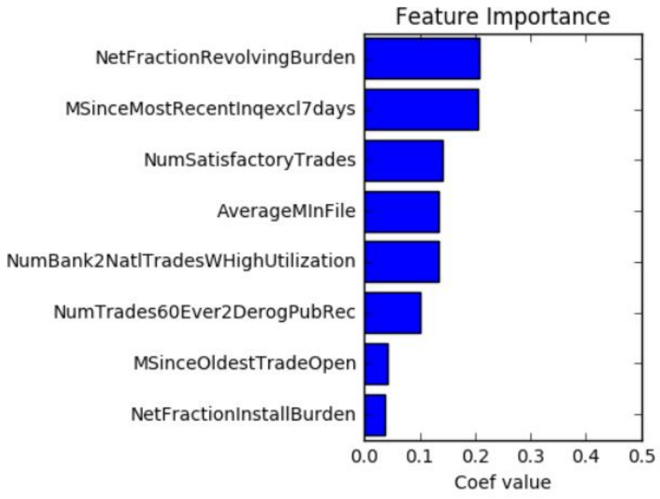


Figure 3: Feature Importance for optimised Gradient Boosting Classifier

Justification

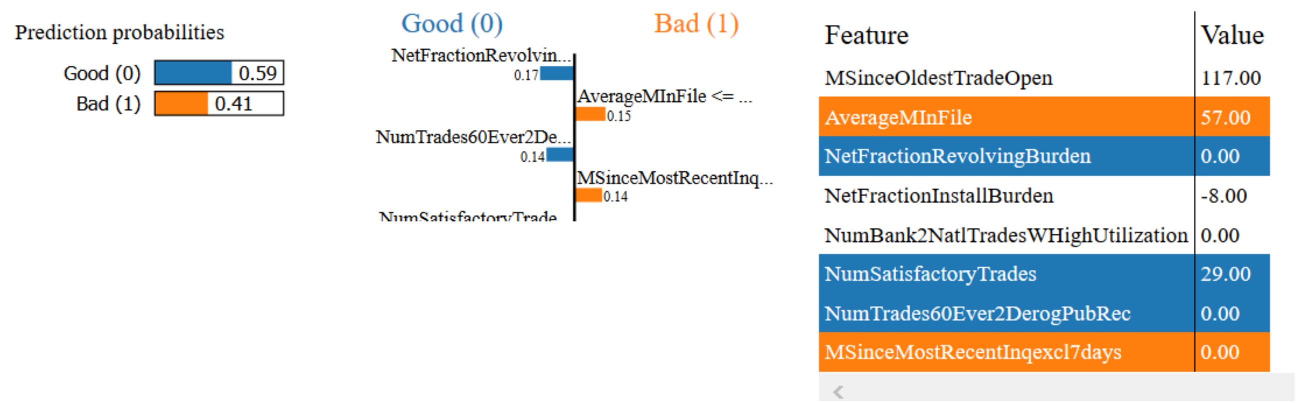
Relative to the benchmark logistic regression model, the Gradient Boosting Classifier has a much higher specificity (74.9% vs 66.8%) which more than offsets the decrease in sensitivity performance (down 2.1% relative to benchmark). This leads to a higher overall Gmean score of 71.2% for the Gradient Boosting Classifier. With the usage of feature importance and LIME, this classifier can be interpreted both globally and locally and hence there is no loss of explanation relative to the logistic regression model. Therefore the Gradient Boosting Classifier has been chosen as the final solution.

V. Conclusion

Free-Form Visualization

The 4 possible outcomes for predicting RiskPerformance will be investigated in this section using the LIME tool.

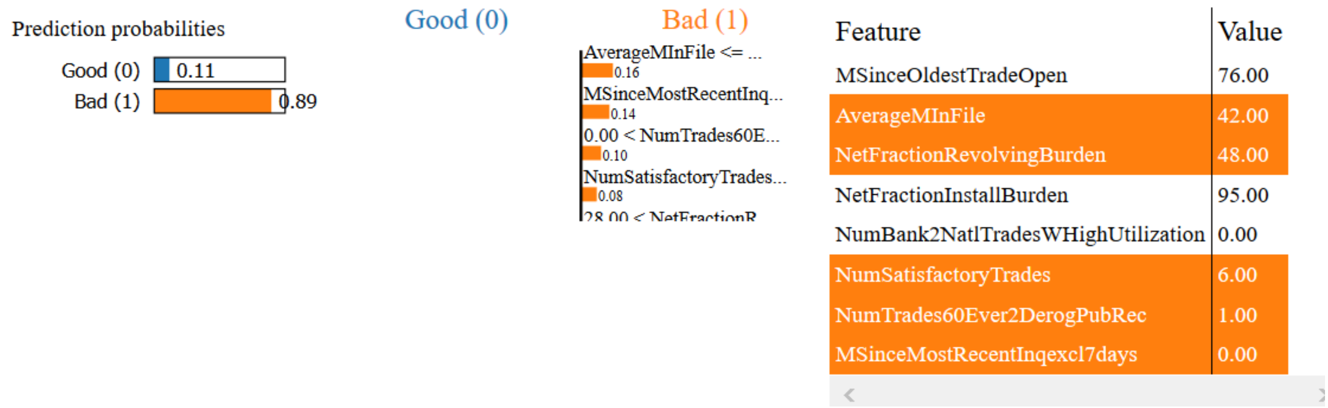
Correct Classification of a Good Loan



Feature	Value
MSinceOldestTradeOpen	117.00
AverageMInFile	57.00
NetFractionRevolvingBurden	0.00
NetFractionInstallBurden	-8.00
NumBank2NatlTradesWHighUtilization	0.00
NumSatisfactoryTrades	29.00
NumTrades60Ever2DerogPubRec	0.00
MSinceMostRecentInqexcl7days	0.00

With no revolving debt and no history of delinquency, the loan is predicted as good despite the borrower's credit history being lower than average and his recent inquiry.

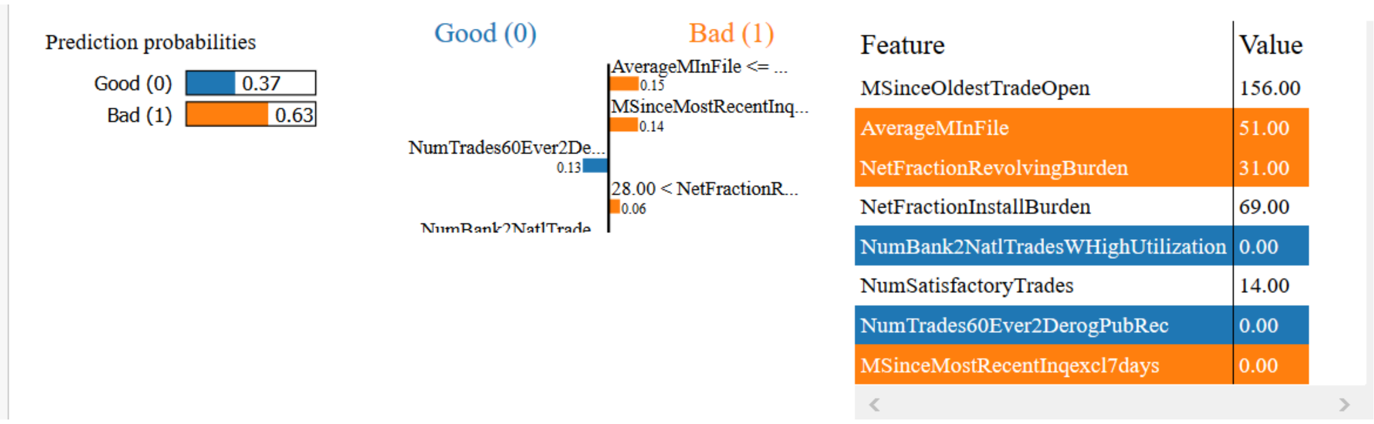
Correct Classification of a Bad Loan



Feature	Value
MSinceOldestTradeOpen	76.00
AverageMInFile	42.00
NetFractionRevolvingBurden	48.00
NetFractionInstallBurden	95.00
NumBank2NatlTradesWHighUtilization	0.00
NumSatisfactoryTrades	6.00
NumTrades60Ever2DerogPubRec	1.00
MSinceMostRecentInqexcl7days	0.00

This borrower has a shorter than average credit history, has a high amount of revolving debt utilisation, missed a payment by at least 60 days and is actively making inquiries for new credit. The model correctly flags this as a Bad loan.

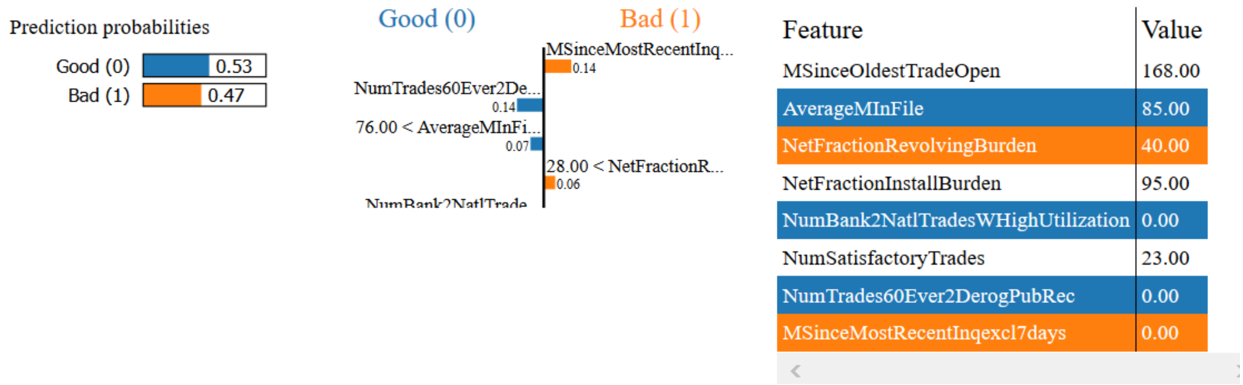
Incorrect Classification of a Good Loan as Bad



Feature	Value
MSinceOldestTradeOpen	156.00
AverageMInFile	51.00
NetFractionRevolvingBurden	31.00
NetFractionInstallBurden	69.00
NumBank2NatlTradesWHighUtilization	0.00
NumSatisfactoryTrades	14.00
NumTrades60Ever2DerogPubRec	0.00
MSinceMostRecentInqexcl7days	0.00

This loan is incorrectly classified as bad as the model picks up on the relatively short credit history of the borrower, his active search for credit and his average revolving burden. In fact, the loan is good, and the borrower has never been more than 60 days late on any payment in the past.

Incorrect Classification of a Bad Loan as Good



Feature	Value
MSinceOldestTradeOpen	168.00
AverageMInFile	85.00
NetFractionRevolvingBurden	40.00
NetFractionInstallBurden	95.00
NumBank2NatlTradesWHighUtilization	0.00
NumSatisfactoryTrades	23.00
NumTrades60Ever2DerogPubRec	0.00
MSinceMostRecentInqexcl7days	0.00

This loan is incorrectly classified as good as the model gives a lot of importance to the fact that the borrower has never been more than 60 days late and has a longer than average credit history. In fact, this loan should be classified as bad as the borrower has a high amount of debt and is actively seeking new credit.

Reflection

The key project steps for this capstone were:

- **Selecting a project:** Given my keen interest in finance, I was looking for a relatively clean dataset which I could use to improve my machine learning skills. Both Kaggle and FICO offered such challenges. I picked the FICO one as I was keen to understand the black box model by using an explanation framework; the Kaggle competition focused solely on score maximisation.
- **Exploratory data analysis:** The key challenge here was how to visualise the different features in a meaningful way to derive insight about the data. I spent a good couple of weeks investigating the various features before settling on the final 8 features to be used for training my supervised models.

- **Model training and performance validation:** This was my opportunity to apply the knowledge I had picked up on the nanodegree and I also wanted to improve my python coding skills using functions and classes. Both objectives were met.
- **Explanation framework:** This was part of the project that I was looking forward to the most and dreaded the most as well. I spent a fair bit going through literature on interpretable machine learning. While my preference was to use SHAPLEY, I was unable to get it fully working on my machine and was not sure whether it was as mature as LIME, so I finally opted for the latter. I would also have liked to go for a more ambitious global explanation framework, but the capstone timelines made me take a pragmatic approach by using feature importance.

Overall, I feel that I must revisit the statistics courses I was taught at university and there is also room for improvement in my python coding. Having said that, I thoroughly enjoy the experimentation aspect in machine learning as well as deriving insights from the data. This is certainly not going to be my last project in this domain.

Improvement

Potential improvements include:

- Experimentation with other features in the dataset which were not used in training the current supervised learning models
- Investigate individual misclassifications to detect common patterns and adjust model accordingly
- Use of other supervised learners such as XGBoost which have proven to be very performant in Kaggle competitions
- Experiment with other advanced explanation frameworks for both global and local interpretation (e.g Shapley value explanations, Example based explanations, Global Surrogate Models)
- Use of area under ROC curve as performance metric to adjust the probability thresholds for Good and Bad RiskPerformance

VI. References

- FICO Explainable Machine Learning Challenge. https://community.fico.com/community/xml/pages/overview
- David Gunning, DARPA. Explainable Artificial Intelligence (XAI). 2017.
- Niels Landwehr, Mark Hall, Eibe Frank. Logistic Model Trees. 2004.
- Tianqi Chen, Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. 2016.
- Finale Doshi-Velez, Been Kim. Towards a Rigorous Science of Interpretable Machine Learning. 2017.
- FICO. Understanding FICO Scores. https://www.myfico.com/Downloads/Files/myFICO_UYFS_Booklet.pdf
- Shunpo Chang, Simon Dae-Oong Kim, Genki Kondo. Predicting Default Risk of Lending Club Loans. 2015.
- Junjie Liang. Predicting borrower’s chance of defaulting on credit loans. 2011.
- Marie-Laure Charpignon, Enguerrand Horel, Flora Tixier. Prediction of Consumer Credit Risk. 2014.
- Yang Liu. The evaluation of classification models for credit scoring. 2002.
- Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, Finale Doshi-Velez. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. 2018.
- Marco Ribeiro, Sameer Singh, Carlos Guestrin. “Why Should I Trust You?” Explaining the Predictions of Any Classifier. 2016.
- Christoph Molnar. Interpretable Machine Learning A Guide for making black box models interpretable. 2018.
- Scott Lundberg, Su-In Lee. A Unified Approach to Interpreting Model Predictions. 2017.
- Marco Tulio Ribeiro. LIME. https://homes.cs.washington.edu/~marcotcr/blog/lime/