Credit Risk Models with Machine Learning Methods

Tan Wei Quan

Agenda

1. What is Credit Risk?
2. What is Expected loss and its components?
3. Why is EL important?
4. Problem statement
5. Methodology
6. Exploratory Data Analysis
7. PD
8. LGD
9. EAD
10. Conclusion

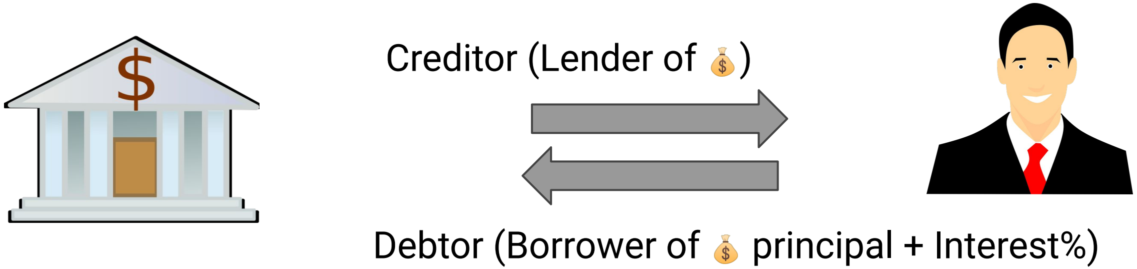
Problem statement

- We are an external credit risk consultancy providing services to Lending Club.
- The aim is to improve credit scoring models processes to assess borrower credit worthiness using ML models
- . Our points of focus will be to:

1. Analyze top 5 features that affect PD
2. Developing a model to predict PD
3. Developing a model to predict LGD
4. Developing a model to predict EAD

What is credit risk?

- Credit risk is the possibility of loss resulting from a borrower’s failure to repay a loan or meet contractual obligation. e.g Credit cards with credit limit or home ownership loan



Creditor (Lender of 💰)

Debtor (Borrower of 💰 principal + Interest%)

Why is EL important?

- Under the Basel 3 Framework, minimum capital requirements under Internal Rating Based (IRB) Approach requires the calculation/reporting of expected loss **(EL).**
- Used by FIs to monitor exposures to customers/credit worthiness.

What is Expected Loss and its components?

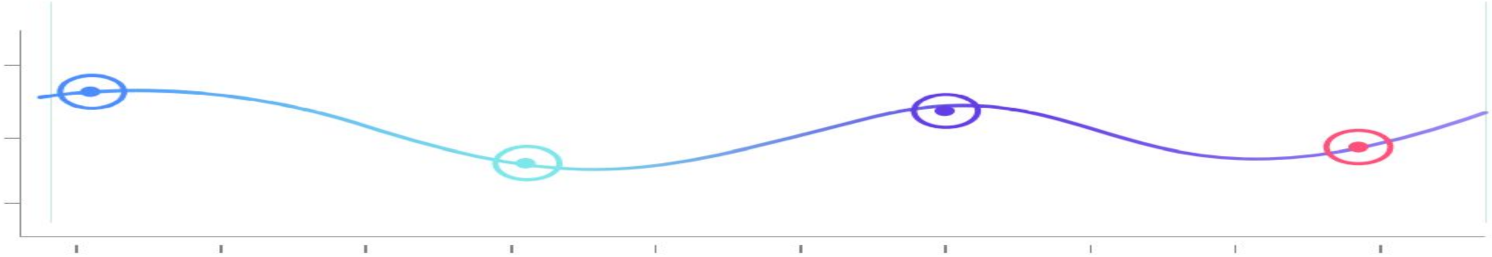
- Expected Loss is the amount a lender stands to lose by lending to a borrower.
- It is calculated by **PD** X **LDG** X **EAD**

(PD) Probability of default represents the chance that borrower defaults

(LGD) Loss given default. Every loan is given at margin & lien. When default occurs, part of the loan can be recovered. LGD represents the proportion of exposure less recoveries.

(EAD) Exposure at default is the amount that lender will lose when borrower defaults

Methodology



Data Gathering
-Source:Lending data ‘07-’15
-EDA (emphasis given on interest rate,term and loan amount)
- PCA

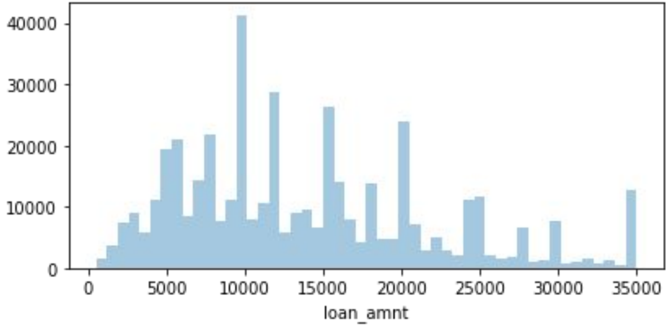
Data cleansing
- Handling of missing data/outliers
-SMOTE
-Feature Engineering

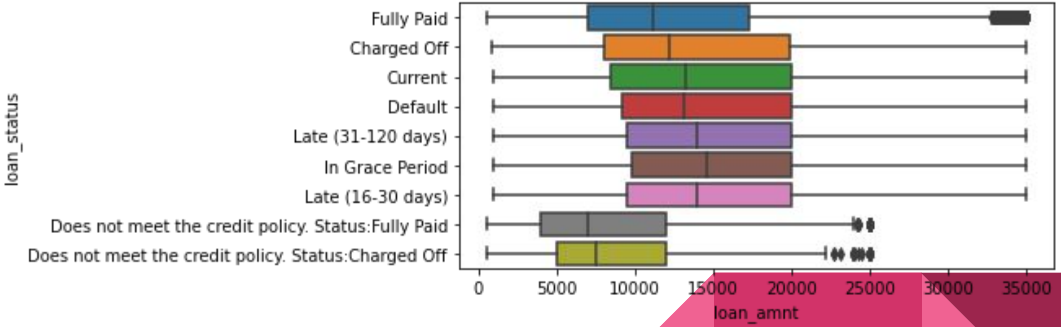
Modelling
- Defining a baseline
- Train, test and refine our current model

Evaluation
- Recommend Features to look for in managing credit risk

Exploratory Data Analysis - Loan amount

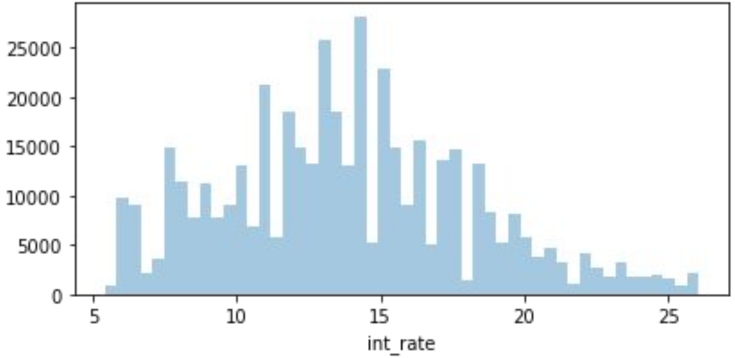
1. We notice that most of the loans are ranging from 10k to 15k.
2. Also, there are seems to be a trend of loans becoming charged off/default to default increases as the loan amount increases from 10k onwards

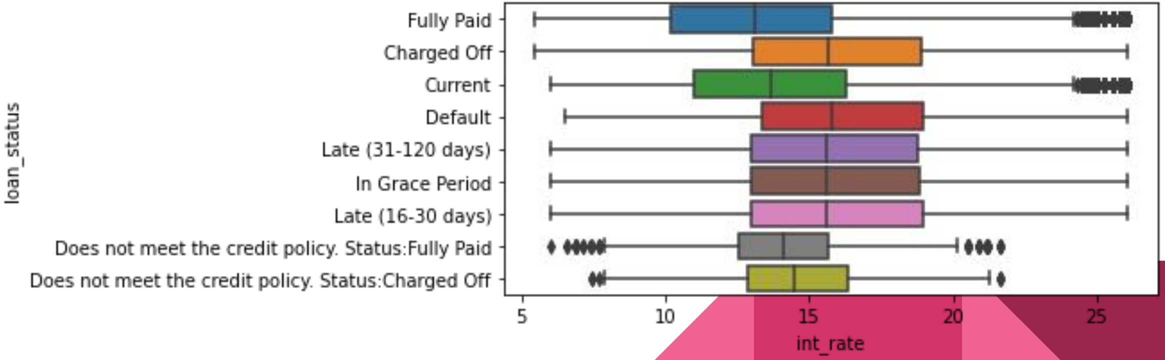




Exploratory Data Analysis - Interest rate

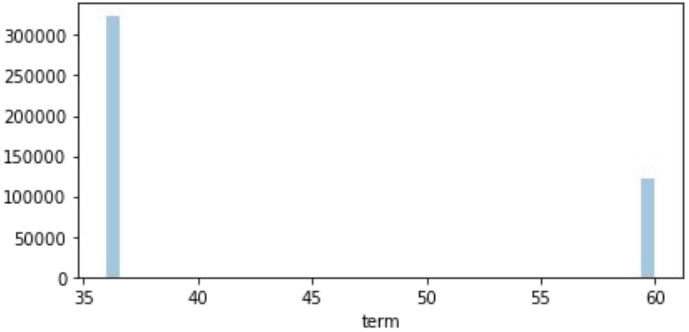
1. We see that most of the loans have relatively low interest rates with a skew of the histogram to the left.
2. Similarly, seems like a trend where loan becoming default/charged off when interest rate increases

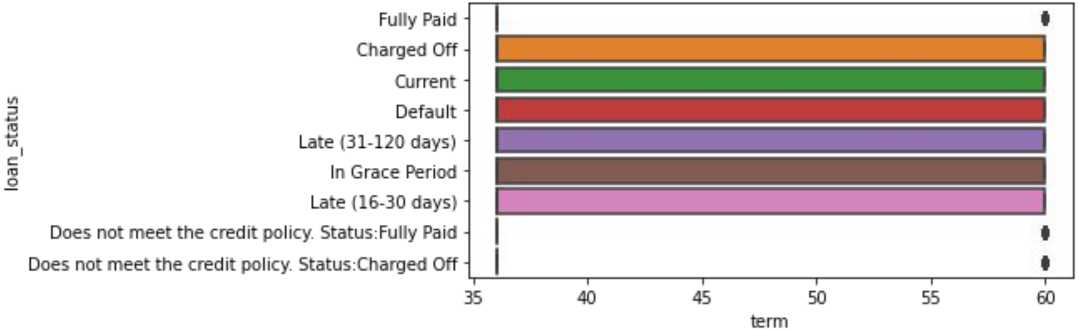




Exploratory Data Analysis - Term

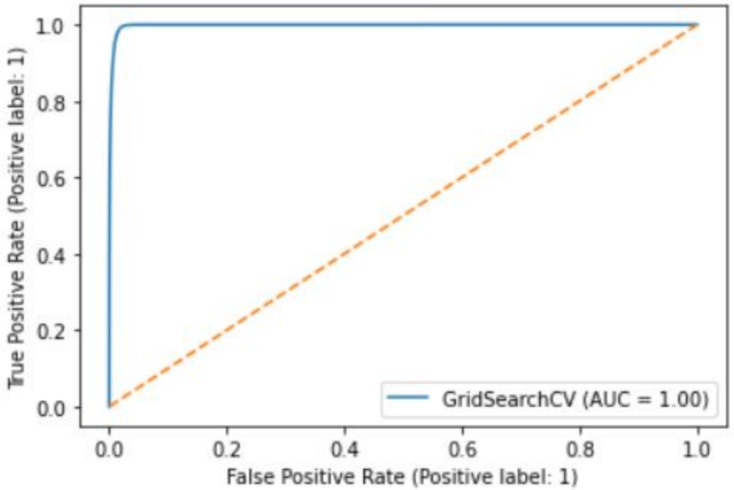
1. Large proportion of fully paid loans are 36 months.
2. For loans above 36 months are mostly in default or charged off.

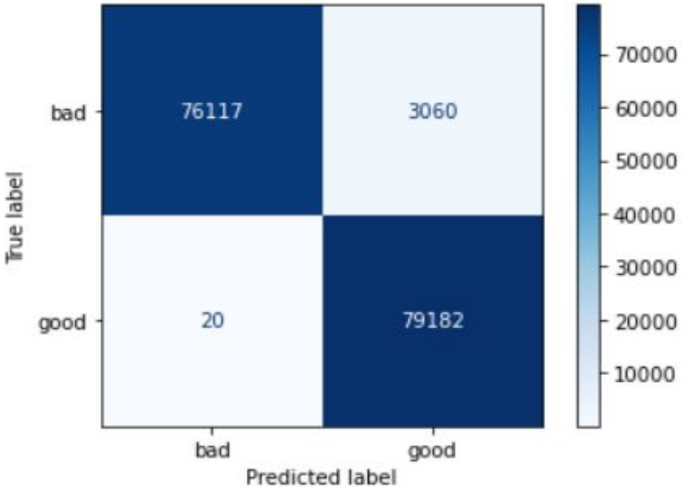




PD models

We will use the Logistic Regression and Random Forest models with hyper parameter Gridsearch to find the best model to predict our probability of default per borrower. Random Forest model yielded the f1 score of 0.98 where the model is able to predict correctly 98% of the time.





LGD

We will use the Linear Regression,Random Forest and AdaBoost,GradientBosst models with hyper parameter Gridsearch to find the best model to predict our recovery rate (recoveries/ funded amount). Random Forest model yielded the best score of 98%

Model	Hyperparameters	score
Linear Regression	Lasso,Ridge Penalties	0.70
Random Forest	'n_estimators'	0.98
Voting Classifier(AdaBoost,GradientBoost)	'ada__n_estimators', 'gb__n_estimators'	AdaBoost 0.84

EAD

We will use the Linear Regression,Support Vector Machines and Neural Networks with hyper parameter Gridsearch to find the best model to predict credit conversion rate(ratio of the difference of the amount used at the moment of default to the total funded amount). EAD is then obtained by multiplying CCF with undrawn amount. Neural Networks yielded the best score 99%.

Model	Hyperparameters	score
Linear Regression	Lasso,Ridge Penalties	0.75
Support Vector Machines	'n_estimators'	0.93
Neural Networks	activation='relu'	0.99.

Conclusion

1. The best machine learning models include ensemble models such as random forests as Neural Network models. Now that we have the 3 components obtained, we will be able to obtain the EL by multiplying in formula **PD** X **LDG** X **EAD.**
2. Research papers on these subjects might have recommended these models usage for specific component calculations. However, the limitations of these models and effectiveness are well documented.

Given more time and processing power, I would look to improve models by adding hyperparameters during gridsearch and also try other dimensions reduction methods. Most research papers cite models with each component to be highly sensitive to a certain number of features

Components	Model	scores
PD	Random Forest	0.98
EAD	Random Forest	0.99
LGD	Neural Networks/SVM	0.93

Recommendations

The top 5 features that affects probability of default are the following:

1. Interest rates
2. Loan repayment in Installment
3. Loan amount
4. Term of loan
5. Debt to income ratio

References:

PD

https://www.moodysanalytics.com/risk-perspectives-magazine/managing-disruption/spotlight/machine-learning-challenges-lessons-and-opportunities-in-credit-risk-modeling

LGD

FULLTEXT01.pdf (diva-portal.org)

SMOTE

https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

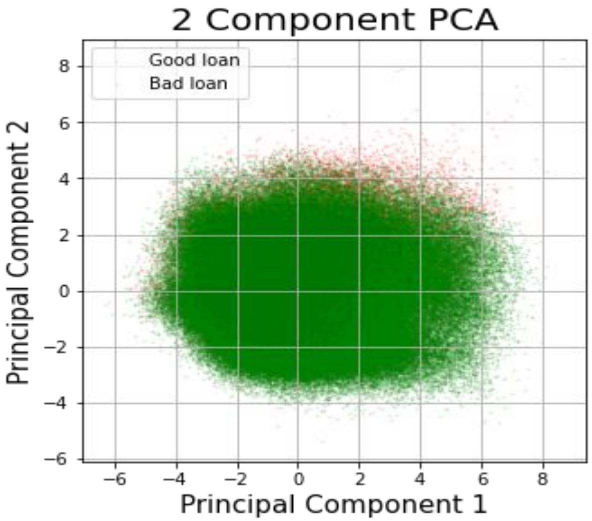
EAD/PD

Finalyse: Machine Learning in Risk Management

Sun-EnsembleLearningCreditRiskModeling-Oct2014.pdf (sas.com)

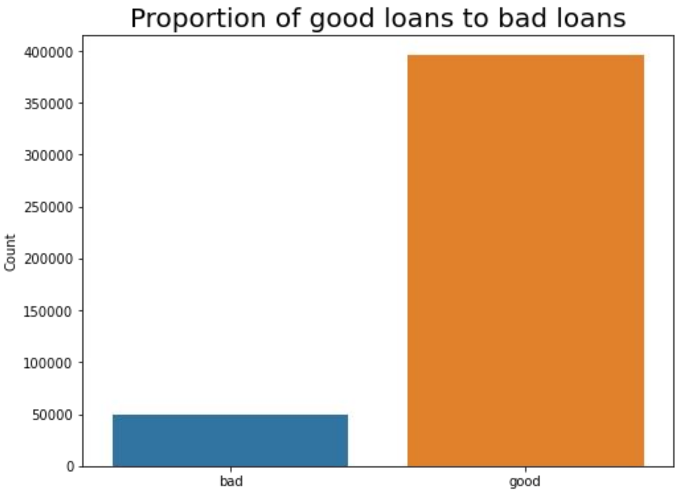
Principal Component Analysis

Using PCA to check variability in my data with principal components, data showed an explained variance of 0.1. Hence, there does not appear to be a standout separation of classes/features.



Class Imbalance

There is a far greater proportion of good loans to bad loans. This will need to be handled during our classification when calculating PD. We have used SMOTE method(oversampling of minority class) to maintain the labelled data balance.


Proportion of good loans to bad loans
Count
bad
good