Rakamin Academy

Virtual Internship Experience

id/x partners

End to End Solution
Credit Risk Scoring Prediction

By: Trianto Haryo Nugroho

Abstract

“This research is a project of a lending company. This study aims to build a model that can predict credit risk using datasets provided by the company. The author makes 5 models of credit rating classifier with the best model showing 98% accuracy. Each model is cross-validated 5 times. The results of this study can be used to measure the credit risk score of someone who applies for credit at lending companies”

Outline

- Background
- Data Introduction
- Target Label Exploration Analysis
- Machine Learning Model
- Webapp Deployment Example
- Conclusion

Background

Credit risk management is the practice of mitigating losses by understanding a bank's capital adequacy and loan loss reserves at any given time. A process that becomes a challenge for financial institutions.

One of the technological advances that is currently a trend is machine learning. Machine learning is considered to be one of the supporters of technological progress for all aspects, especially in data processing.

With machine learning, we can provide technological solutions for lending companies by building models that can predict credit risk.

Lending Company Dataset

- All data are of individual loan application type
- Data taken from 2007 to 2014
- In the dataset there are 466,285 rows of data and 75 columns

Label Target

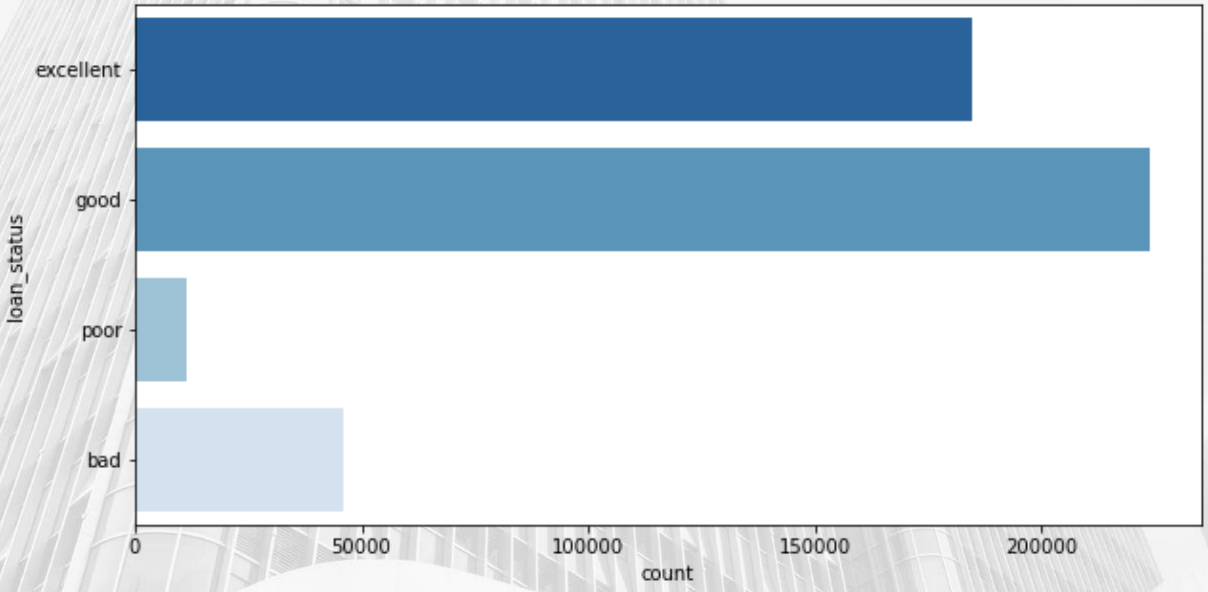
To get insight from the dataset, the first thing to do is to determine the target label. The loan_status label target is simplified into the following 4 classifications:

- Excellent means that the person has a credit history who have been fully paid and have no problems
- Good means that the person is doing credit and has never had a problem
- Poor means that the person has a history of being late pay
- Bad means that the person has a history of failure pay

Next we can draw insight by looking at some features with target labels

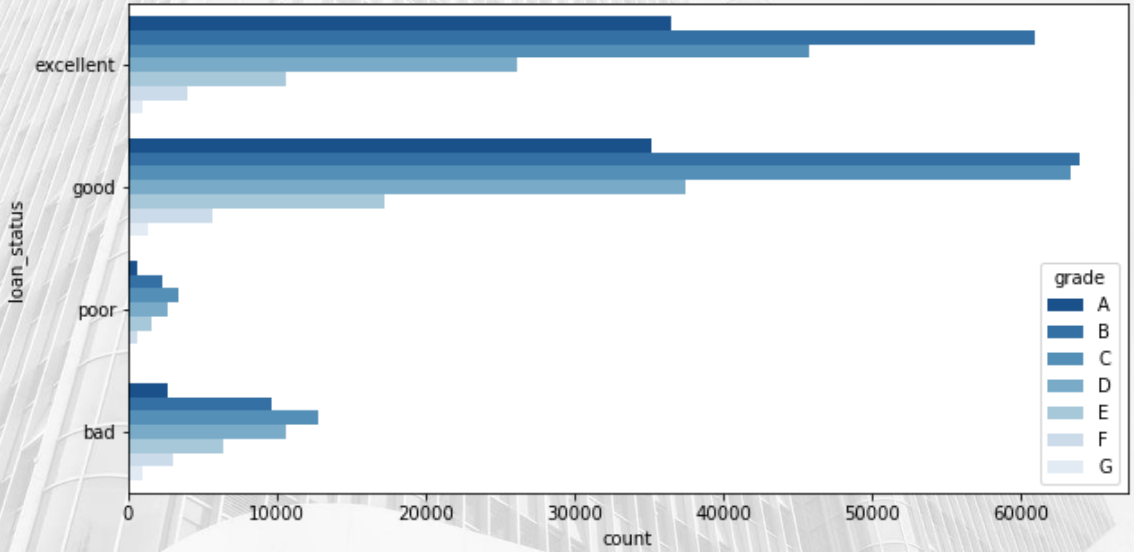
Label Target Distribution

The results obtained with Target Label with distributions like the following, data imbalance due to poor classification is very unequal with excellent and good



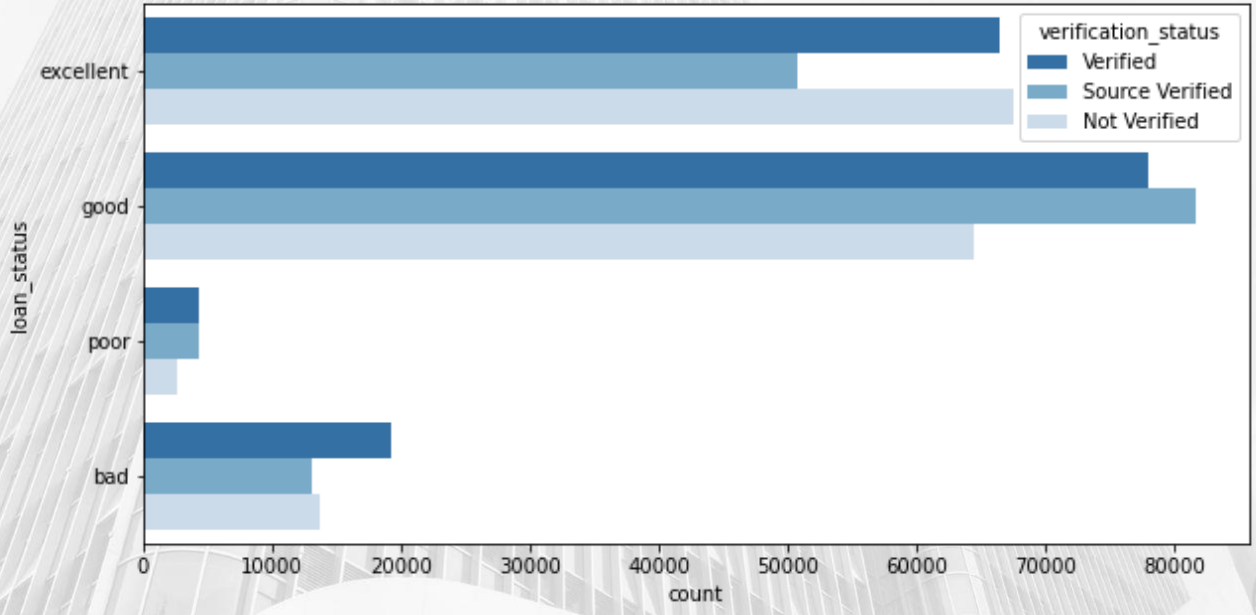
Label Target VS Grade Analysis

When compared to grades, there are some with Grade A still classified as poor and bad, and vice versa. There are several Grade G which are classified into good and excellent categories. Grading does not necessarily indicate that the credit rating will be good too.



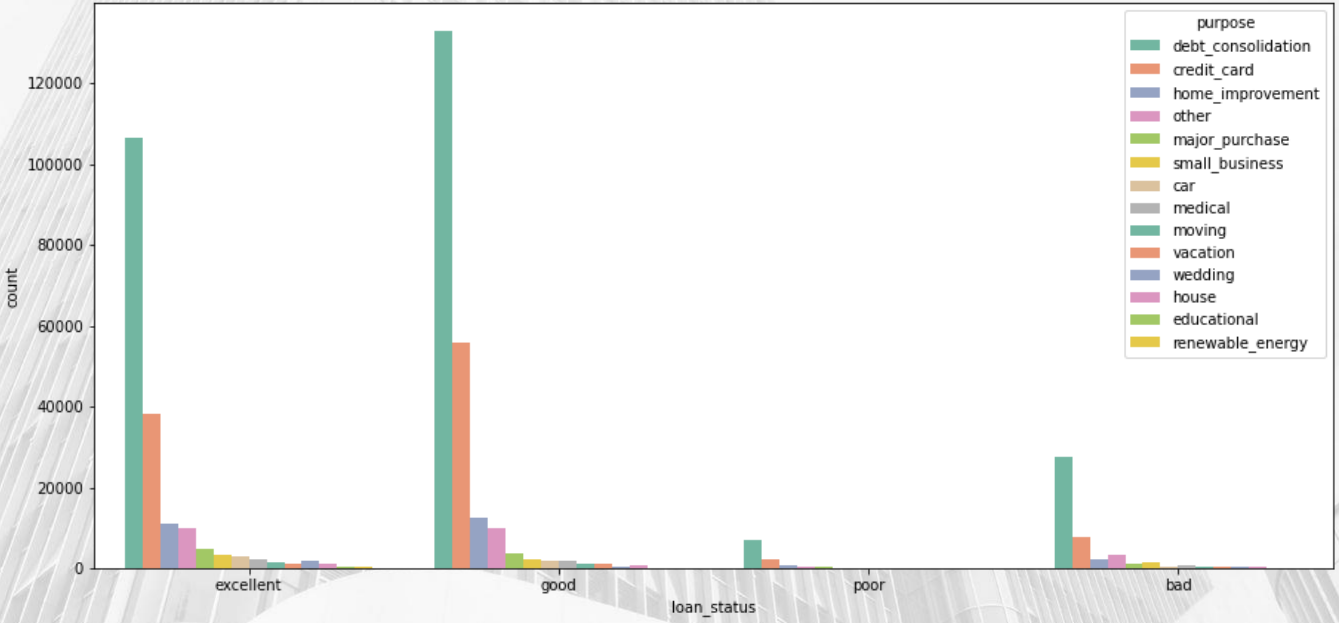
Target Label VS Verification Analysis

When compared to verification status, even though there are many borrowers who do not have Verified status, it does not guarantee that the Credit Rating will be good either.



Target Label VS Loan Purpose Analysis

The purpose of most borrowing is debt consolidation and then credit cards. The distribution of loan objectives and each classification of target labels are quite similar, so that no single loan objective is categorized as bad credit risk.



Machine Learning Model

From the dataset, outlier handling, missing value handling and feature selection were carried out. More is explained in the Python notebook Chapters IV, V, and VI. Then 5 models were made with the results of the modeling as follows with an accuracy score that has been cross-validated.

The best model is XGBoost

Decision Tree Model	
Excellent F1 Score	0.98
Good F1 Score	0.98
Poor F1 Score	0.54
Bad F1 Score	0.92
Accuracy Score	96.4%

Random Forest Model	
Excellent F1 Score	0.98
Good F1 Score	0.99
Poor F1 Score	0.64
Bad F1 Score	0.92
Accuracy Score	97.3%

Naïve Bayes Model	
Excellent F1 Score	0.95
Good F1 Score	0.99
Poor F1 Score	0.64
Bad F1 Score	0.68
Accuracy Score	94.2%

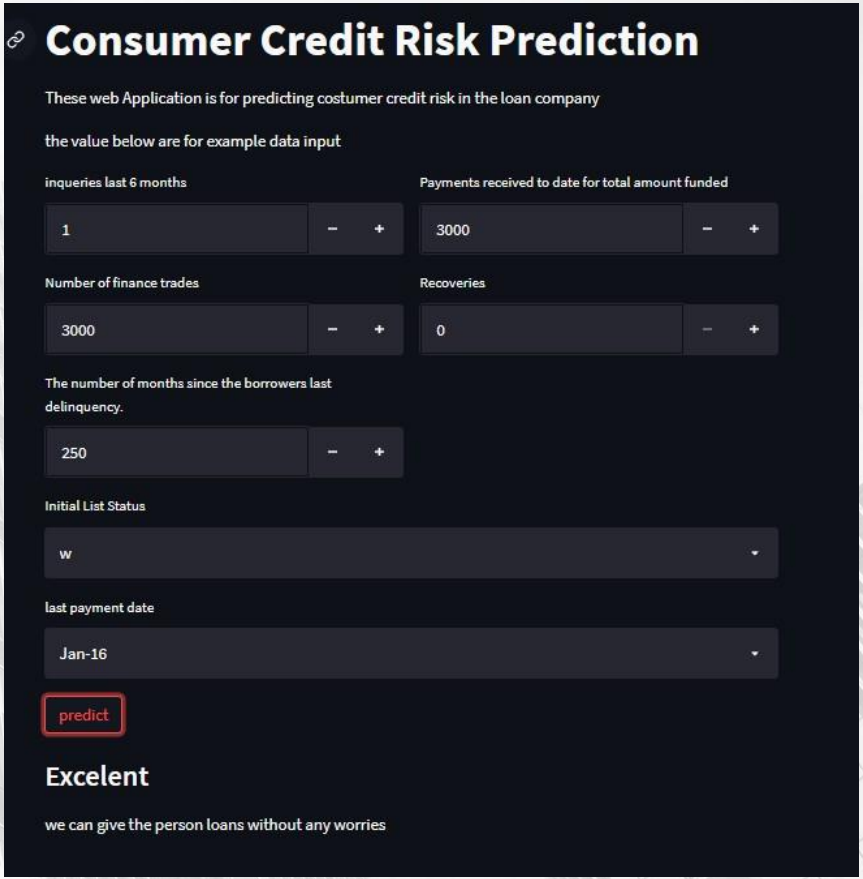
KNN Model	
Excellent F1 Score	0.97
Good F1 Score	0.98
Poor F1 Score	0.62
Bad F1 Score	0.87
Accuracy Score	96.0%

XGBoost Model	
Excellent F1 Score	0.99
Good F1 Score	0.99
Poor F1 Score	0.66
Bad F1 Score	0.95
Accuracy Score	98.0%

Model Deployment

The best models are deployed to Heroku using the python library Streamlit to build the Webapp.

The form is filled in according to the data from the borrower, then predict to get the credit risk scoring results. The model can also be deployed to Cloud Web Services.



Conclusion

For the target label with grade, verification status and loan purpose, they are not very related to the target label, respectively. because for the distribution of grade, verification status and loan purpose, each has a similar distribution for each classification of the target label.

The results of each model from accuracy cross validation are Decision Tree with 96% accuracy, Random Forest with 97% accuracy, KNN with 96% accuracy, Naive Bayes with 94% accuracy, XGBoost with 98% accuracy So the best model is the XGBoost model so that model is selected as the model for deployment.

The results of this model can be used to help a credit risk analyst to determine whether a borrower can borrow with the least risk by lenders. By using a model that has an accuracy of 98%, a credit risk analyst can find it easier and faster to determine whether a borrower is good to get a loan or not from a lender company.

Thank You!

LinkedIn	https://www.linkedin.com/in/trianto-haryo-nugroho/
GitHub	https://github.com/triantonugroho