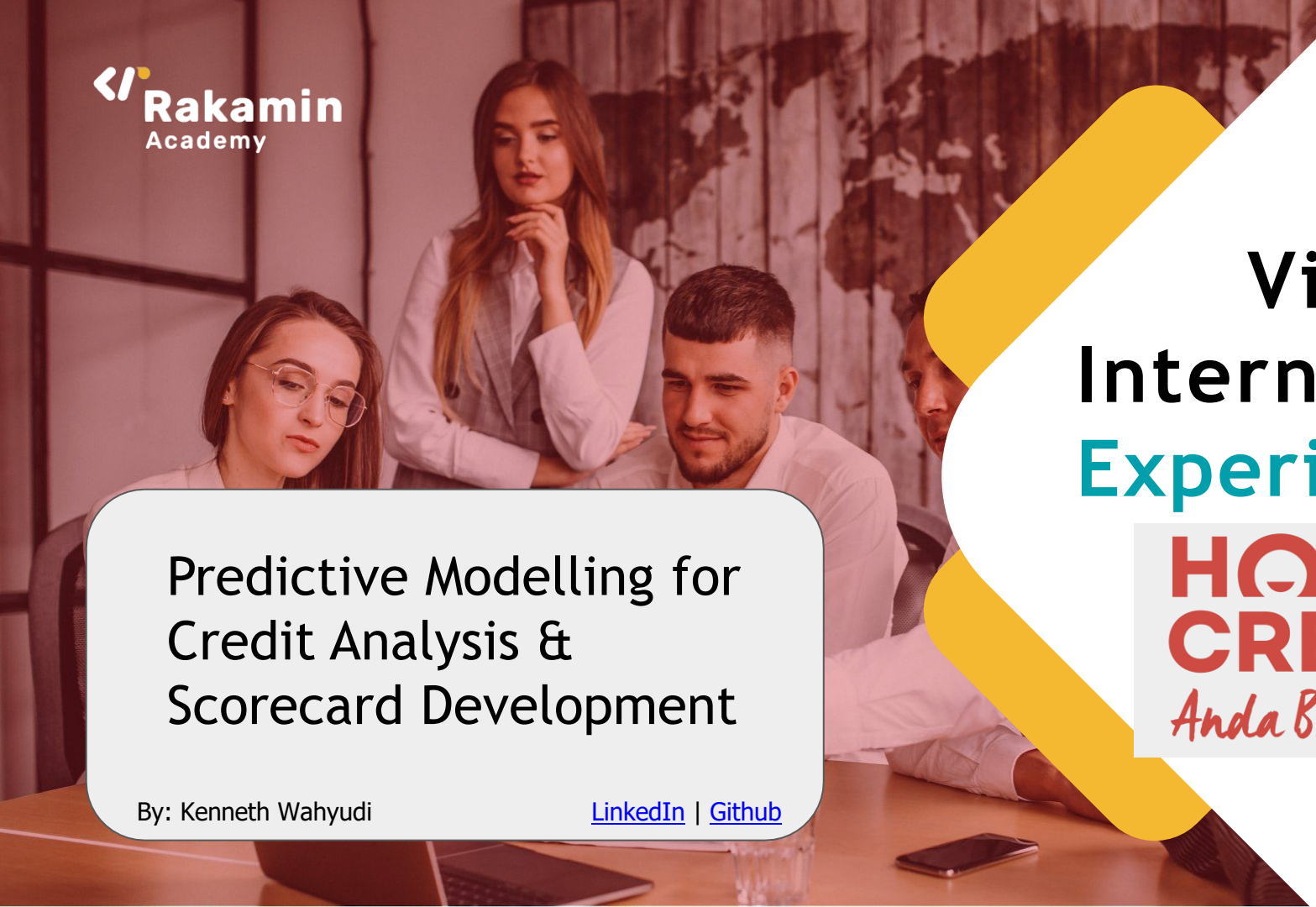


Rakamin Academy

Virtual Internship Experience

HOME CREDIT Anda Bisa!

Predictive Modelling for Credit Analysis & Scorecard Development

By: Kenneth Wahyudi

LinkedIn | Github

Table of Contents

The dataset used in this project can be accessed in the following link.

The complete code in .ipynb file format can be found in the following link.

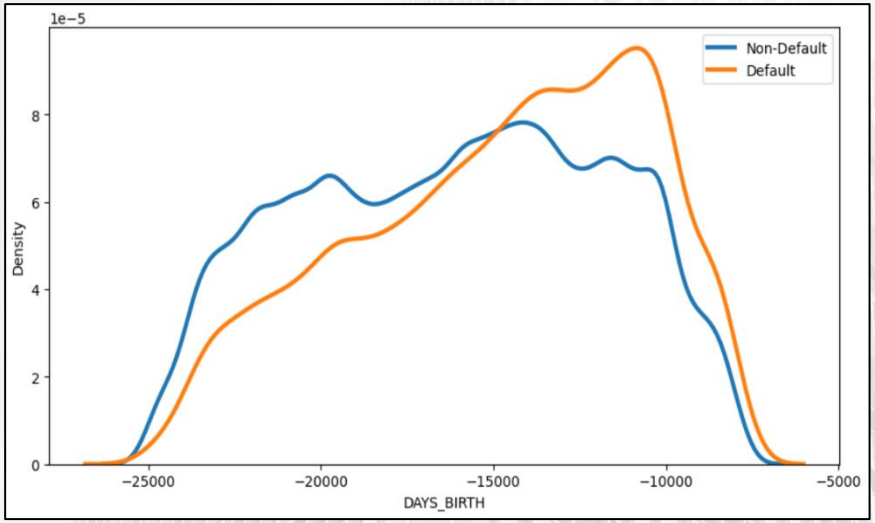
Problem Overview

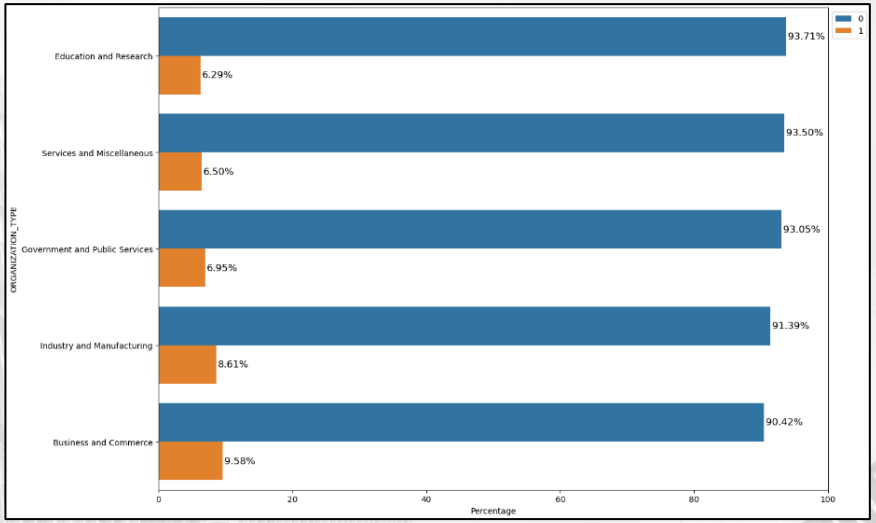
Home Credit is currently leveraging various statistical methods and Machine Learning techniques to predict credit scores. However, the challenge lies in unlocking the full potential of our data. By doing so, we can ensure that creditworthy customers are not denied loans, and loans are provided with terms that inspire customer success. The key is to optimize the use of data through the application of at least two Machine Learning models, with Logistic Regression being one of them. The evaluation will gauge the depth of your analytical understanding. Our task is to create a presentation slide showcasing end-to-end modeling analysis and providing actionable business recommendations for a successful lending strategy.

Dataset Overview & Preprocessing

The dataset contains over 300k rows and over 200 columns after all the tables have been properly merged together. With the numerical features mostly consist of multimodal distribution, and the categorical features mostly have numerous amount of unique values. For the feature selection, we used the information value (IV) as a guideline. With that, there were only 47 features deemed usable after encoding. And because we're using logistic regression as one of our models, we scale the features using Min Max Scaler. And to handle the imbalance, we used the Random Under Sampler. Since using the oversampling techniques will result in very exhaustive and expensive computer processes.

Data Visualization & Business Insight





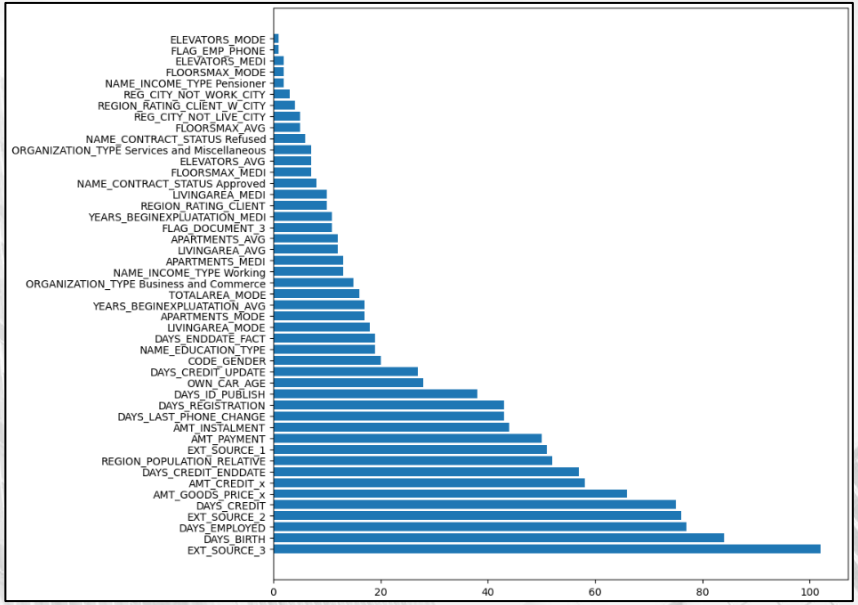
Two of the most interesting insights are age and organization type. The older the customer is, the more reliable they are in paying their loan properly (left chart). And the Business & Commerce organization type is the least reliable with a default rate of 9.58%. That is 52.3% higher than the Education & Research organization type with a 6.29% default rate (Right Chart).

Model Development & Evaluation

After trying different models, the models that are finally used are the XGBoost Classifier and Logistic Regression (scorecard model). After the hyperparameter tunings, the XGBoost model's metrics are 0.7 in recall and 0.75 in ROC AUC, and the Logistic Regression model's metrics are 0.7 in recall and 0.74 in ROC AUC.

The feature importances for the XGBoost model is shown on the chart to the right.

As for the scorecard, it can be accessed in the following link.



Business Simulation

Before the modelling, when we accepted all applications, the default rate was around 8% out of the whole application accepted. However, with the predictive model, out of all the applications accepted, there would be less than 5% (3.77%) of defaults. For further simulation, we need to make assumptions. Let's say, that on average, every loan is 1,000 dollars in amount, and the interest that we are going to get from that is 100 dollars. And let's say, that on average, when a customer defaults, we lose 50% or half of the amount that we loaned to them & its interest as well. Now that we get that out of the way, it's time to simulate.

Business Simulation

Before the modelling, there are 307511 customers, of which 24825 defaulted. That means that out of all the loan that we gave (307511 x 1000) = 307,511,000 dollars, we receive back (282686 x 1100) + (24825 x 550) = 324,608,350 dollars which is a 5.56% profit.

Now after the modelling, looking at the confusion matrix in the previous section, out of 30752 applications, we only gave out loans to 20605 customers, of which 777 defaulted. That means that out of all the loan that we gave (20605 x 1000) = 20,605,000 dollars, we receive back (19828 x 1100) + (777 x 550) = 22,238,150 dollars which is a 7.93% profit.

Business Simulation

We also need to address the fact that there are opportunity losses, since we are maximizing the recall score. If we do the first scenario (giving loans to all customers that applied) on the test dataset, than out of all the loan that we gave (30752 x 1000) = 30,752,000 dollars, we receive back ((19828 + 8441) x 1100) + ((777 + 1706) x 550) = 32,461,550 dollars. That's a profit ratio of 5.56%, and a total profit of 1,709,550 dollars, whilst the second scenario previously had a total profit of 1,633,150 dollars.

Therefore in conclusion, there is a 42.62% increase in profit margins after implementing the predictive model. Without losing too much on opportunity loss.

Business Recommendation

Despite all that, the models still have a lot of room for improvements, especially metrics wise. Therefore, we also need to look at the insights when accepting (or even targeting) customers for their applications. The top features are external sources score, age, and length of employment. Customers who didn't default on their loan usually have a high external sources score, are of mature (or even senior) age, and have been in the working force longer than most. Although serving only as guidelines, if we can target such customers to apply for a loan (through marketing or discounts), then it should help the predictive model and credit scorecard in achieving its collective goal of minimizing default rates and maximizing profits.