Machine Learning Individual Project

Manyi (Mandy) Luo

1003799419

January 4, 2022

1 Individual Business Write Up

To explain the business value of the project's main findings and how findings can be applied, I would like to expand it further regarding two aspects: strategy changes and the deployment of models. The project aims to investigate home credit default through applying three machine learning models, Random Forest, Boosted Tree and Logistic Regression, and accessing their predictive power respectively. This helps us to identify clients with the potential credit risk of payment difficulties, which is defined as late payment time on at least one of the installments of the loan, rather than those who struggle to get loans due to insufficient or non-existent credit histories. To make the models able to generalize, a variety of alternative data are included, for example, telco and transactional information to predict the clients' repayment abilities and to ensure that only incapable repayment are captured. The main strategic change is that the machine learning approach includes more essential information compared to the traditional approach that accesses credit status according to past credit history. As a result, we can prevent qualified loans from being rejected due to the lack of credit history. We can also conduct feature engineering and include more relevant features to access the client. From the aspect of deployment, machine learning techniques can potentially automate the process, so more manpower and employment costs will be saved. Therefore, the main business value of this project is to increase the accuracy of accessing client credit performance and reduce credit risk exposure by correctly identifying clients with the inability to make payments.

Next, I would like to provide some intuitive explanations for feature engineering, best features and other insights in the model construction process, which are used for logistic regression. First, feature engineering is the process of using existing knowledge within the data set to select and transform the key variables when creating a powerful machine learning model. Through performing feature engineering, it's able to improve the model's predictive performance, reduce computational or data needs, and enhance the interpretability of the modeling results. In this project, some irrelevant features or features having constant values are dropped during the data cleaning stage. Also, feature transformation and merging are conducted to guarantee the significance of features in the preliminary process. We then perform feature engineering to transform data and to obtain the Weight of Evidence (WOE), which is calculated based on the distribution of accounts without payment difficulties versus those with payment difficulties. Through reducing the multicollinearity by clustering, the step-wise logistic regression is performed using the selected features with the highest information value from each cluster, illustrating the best predictive power. In the modeling process, the best features are selected based on feature importance and information value. According to the group report, the top four best features are `EXT_SOURCE_1`, `EXT_SOURCE_2`, `EXT_SOURCE_3`, `DAYS_BIRTH`, and `DAYS_EMPLOYED`, which represent the normalized score from external data, clients age, and days before the application the person started current employment. These variables make sense because stable clients with moderate age and healthy employment history will have less trouble in meeting any payment difficulties, which makes the logistic regression model representative.

2 Individual Technical Write Up

In this section, I would like to explain the model validation and approach chosen in this project, as well as how to choose the best model and its hyperparameters. The truth-worthiness is self-contained in the processes. First of all, for the two machine learning models, random forest and gradient boosted tree, the k-fold cross-validation is implemented. As outlined in the group report, this method takes the validation data from the training data set, but it does not participate in

the training. As its name suggests, this method divides the original data set into K partitions, which means that each subset in the k-fold can be used as a validation set, and the remaining k-1 subsets can be used as the training set, which creates the K models. The technique used is to plot the corresponding Operating Characteristic Curve (ROC) for each fold by evaluating the K models separately in the validation set. The curve acts as a performance measurement under different threshold settings and the Area Under Curve (AUC) is used to represent the measure of separability given the probability curve. As a result, this statistics is able to tell the ability of the current model to distinguish between various classes. Note that AUC ranges from 0 to 1, and as it gets closer to 1, this means that the model has an outstanding separability. After measuring the goodness of fit through cross-validation, the optimal hyperparameters of random forest and gradient boosted tree are found through the sequential grid search method. By implementing the grid search, the first step is to assess the impact of each parameter on the given machine learning model performance, so it's possible to select the most influential parameter to conduct optimization that is limited within a given range to ensure efficiency. During the process of the grid search, the target parameter will be picked one by one, while holding all other parameters constant. Note that using grid search cross-validation, each parameter is hyper-tuned, which yield a new parameter that is tuned using the model with the partial set of tuned parameter. As a result, it's possible to choose the target parameter which maximizes the AUC as the optimization result for a specific target parameter.

Moving on to the step-wise logistic regression, the bidirectional elimination is implemented to choose the optimal explanatory variables. In this case, Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) are chosen as target statistics, as they are able to estimate the relative amount of information lost by a given model, and give penalization for more loss respectively. The parameter tuning begins with a model with no explanatory variable, and then it adds more variables that improve the model the most by reducing AIC or BIC statistics. This process is done one variable at a time and it stops when no additional variables can yield any improvements. Note that due to variable correlations, it might delete a selected variable in some special cases. When the model complexity gradually increases and there are no more under-fitting issues, the optimal

is being reached, so the lowest generalization error and the highest model accuracy are obtained. After reaching this optimal point, the generalization error increases again, which leads to undesired over-fitting issues. Therefore, for hyperparameter tuning, the model attempts to adjust the complexity by making the generalization error to reach its minimum point. As a result, a trustworthy model is derived with both desirable accuracy and complexity.

In conclusion, according to the group report, it was found that the performance of the tuned random forest model is slightly better than the logistic regression model with the engineered features, and the gradient boosted trees model only performed better for some metrics. Generally speaking, the baseline model and logistic regression model with the default data set performed the worst, whereas the random forest model and the logistic regression model with the feature engineered data set performed well. Overall, all three models were able to make accurate predictions on the likelihood of a client defaulting on a loan, and the models appear to generalize well.

Comparing these models, random forest reduces the influence of outlier, as it selects a proportion of the data set to establish multiple decision trees, so the prediction results are obtained by a range of decision trees rather than single, which will reduce the impact of a single outlier. It also reduces the effect of overfitting, as it only utilizes partial features to construct several decision trees through sampling with replacement, which means that the data within a single decision tree has less proportion, reducing the occurrence of overfitting. Note that random forest can handle large data sets as well since the model has generalization ability after consistent sampling. The existence of missing values also has less effect on the random forest model. In conclusion, this provides the reason for the random forest being the best performance model. Note that the random forest model after hyperparameter tuning is better than the baseline model since its parameters are optimized to improve the model performance.

I would also like to explain the gaps in the performance of models from traditional logistic regression model as well as machine learning models. Generally speaking, machine learning models, random

forest and gradient boosted tree, are non-linear models, which will outperform linear model when dealing with unstructured data set. Once notice corresponds to this trait and I observed that logistic regression performs better after feature engineering, especially when the data becomes more structured, meaning that there is less difference between non-linear model and linear model.

If there's more time to improve on the best model, I might consider doing more feature engineering to improve the input data quality, including the method to deal with missing values rather than replacing them with zeros (like considering the mean value), to create more data transformation with devices having better computational power, new variable creations to replace given variables for less redundancy and check correlation matrix to keep make sure all the features are highly correlated with the target dependant variable. Also, for model improvements, it's important to note that including WOE for the machine learning model would improve its predictive power as well and it's possible to conduct more hyperparameter tuning and consider the hyperparameters within a greater range in the optimization stage.

3 Outline Next Steps

Last, I would like to provide some suggestions if I become the new VP for machine learning in credit risk. In my two-years plan, I will outline what kind of data to collect, what kind of modeling techniques to apply, and what products to create and integrate into the overall business strategy. Regarding the data usage, just like the data results presented in this project, I would consider using feature importance, calculating information value, assessing feature correlation to get a comprehensive idea of the type of data that is suitable for collection. What's more, I would take an eye on whether the current data set could be manipulated to have higher quality, for example, the existence of less missing value, the comprehensiveness of available data, the structure of data set that is balanced or not by having similar numbers of $TARGET = 0$ and $TARGET = 1$.

Regarding modeling techniques and applications, it's possible to utilize a more comprehensive data

processing scheme. For example, regardless of extra time consumption, it would be helpful for feature engineering to use an advanced method such as Principle Component Analysis (PCA) in feature selection and to organize data more efficiently by using clustering methods, which can also facilitate the understanding of how the model makes predictions. Additionally, it's always crucial to keep an eye on the data set and continually collect new data as it becomes available, so the model can keep improving throughout time. For modeling, it's possible to incorporate advanced techniques as well, such as AutoML to train high-quality models or consider using a combination of different models.

Regarding the products to create and integrate into the overall business strategy, I would consider creating a new credit product or modeling strategy. If the model can provide a good prediction of default rates, riskier credit products can be potentially deployed. For example, considering releasing a credit card with a higher credit limit might attract new customers to make business loans with the company, and the credit modeling will only accept stable customers who are eligible to make regular payments. Besides that, it's possible to examine the customer's traits and features of non-default and to perform a market campaign about more credit products with these selected customers. By issuing new credit products, the profit generated from selected customers after machine learning model screening would outweigh the product deployment costs and fees related to monitoring and examining the credit product, which is profitable for the company overall.