**BT 4016: Assignment 2: Credit Risk**

Due: 8 November 2019, 11:59 pm

Assignment Worth: 11% of your grade (22 points)

The goal of this assignment is to build reduced form models to predict loan default from a set of loan attributes. In this assignment, we will create models using a lending dataset from Lending club, and test the predictive ability of different algorithms.

A dataset containing complete loan data for all loans issued is available on LumiNUS. It includes the current loan status (Current, Late, Fully Paid, etc.) and a large set of attributes for each customer.

**Please submit a word document with your answers and the jupyter notebook.**

Full model

The dataset on LumiNUS has a mix of continuous and binary attributes for each loan, along with an indicator of the "loan status." Create an indicator for default (1) if the loan was charged off, or late in payment, and 0 otherwise.

First, we will try to use all available information (all attributes in the data set) to predict if a loan will default (1/0). We call this the FULL-MODEL.

Second, you will split the dataset into two parts: 80 % training, and 20 % test. **Apply random\_state = 1234 for reproducibility** (for **all** questions). 1 point will be deducted for those who do not apply this setting.

1. (3 points) Apply three different techniques (logistic regression, tree classification, XGBoost) to all attributes in the dataset (FULL-MODEL) to predict loan default.

Compute the accuracy and MCC of each measure to compare how good each technique is at using these attributes at predicting default, using a test set.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Logistic regression | Tree classification | XGBoost |
| Accuracy |  |  |  |
| MCC |  |  |  |

2. (2 points) Describe the strategy you took in terms of training the data and selecting the appropriate parameters to best avoid overfitting. This should also be applied in the above and subsequent questions for consistency. \*Hint – you may further divide the training data\*

Reduced model

Not all of the attributes will be useful for predicting if a loan will default. The next step is to reduce this number of attributes to get a more parsimonious model, while still being able to predict default well.

1. (2 points) Which loan attributes *do you believe* are the most informative? Please use your **knowledge and intuition** to choose 10 of the attributes available, and explain why you chose these attributes. Let’s call this the REDUCED-MODEL.

1. (2 points) Let’s use the attributes you’ve selected for the reduced model to predict default, and check the performance of the resulting models. Apply the same three techniques (logistic regression, tree classification, and xgboost) to the REDUCED-MODEL (i.e. only using the attributes you selected) to predict loan default. Again, compute their accuracy and MCC.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Logistic regression | Tree classification | XGBoost |
| Accuracy |  |  |  |
| MCC |  |  |  |

Lasso- reduced model

Let’s now try to use a statistical approach to select the variables.

1. (a) (2 points) Calculate the accuracy and MCC of the LASSO approach, and compare it to previous methods. Apply the LASSO to the default response variable (1 or 0), and the entire set of 70 attributes, and constrain the model to have at most ten attributes in it. Find the best lambda. Let's call this the LASSO- MODEL.

Which attributes are those?

Compute the method’s accuracy and MCC using 5-fold cross validation

|  |  |
| --- | --- |
|  | LASSO |
| Accuracy |  |
| MCC |  |

(b) (2 points) Did your intuition about the correct set of attributes in the REDUCED-MODEL match the results that you obtained from the LASSO-MODEL? Discuss.

Understanding Model Performance

6. (1 point) What other metric might you use to compare the performance of the different models? Explain what this metric means and how it might be computed. For this question, you will not have to compute the metric.

7. (2 points) How do the results using the full set of attributes FULL-MODEL and LASSO-model compare to those that you personally selected REDUCED-MODEL to predict if a loan will default? What do you think is more important in this particular example, the set of attributes, or the classification technique? Discuss.

9. **Open ended Tuning Problem**

Your manager has given you a task of tuning a model that could be operationalized for predicting loan defaults. He has asked you to use the given full dataset and has left it up to you to decide how to use it. Your job is therefore to come out with a model that can generalize well, hence overfitting to data is of a concern as the dataset is relatively small.

1. (3 points) Describe your validation and testing procedure, noting down what techniques worked for you and what did not. Tabulate any results across the different tuning attempts and note down the various techniques you tried (e.g. feature engineering, different base models etc) in a separate column.
2. (3 points) Also discuss any insights you may uncovered while tuning the models (what parameters made the most difference for which models, etc)

Explanatory data analysis from various techniques/tools you have learnt in your other modules or beyond the scope of the classes (e.g. stacking with other models, Bayesian optimization) may also used to further identify candidates for feature selection and engineering.

*\*This is an opportunity to show your skills as a data scientist and the problem should be approached as such – describing what you did to arrive at the model is as important as producing the best model\**