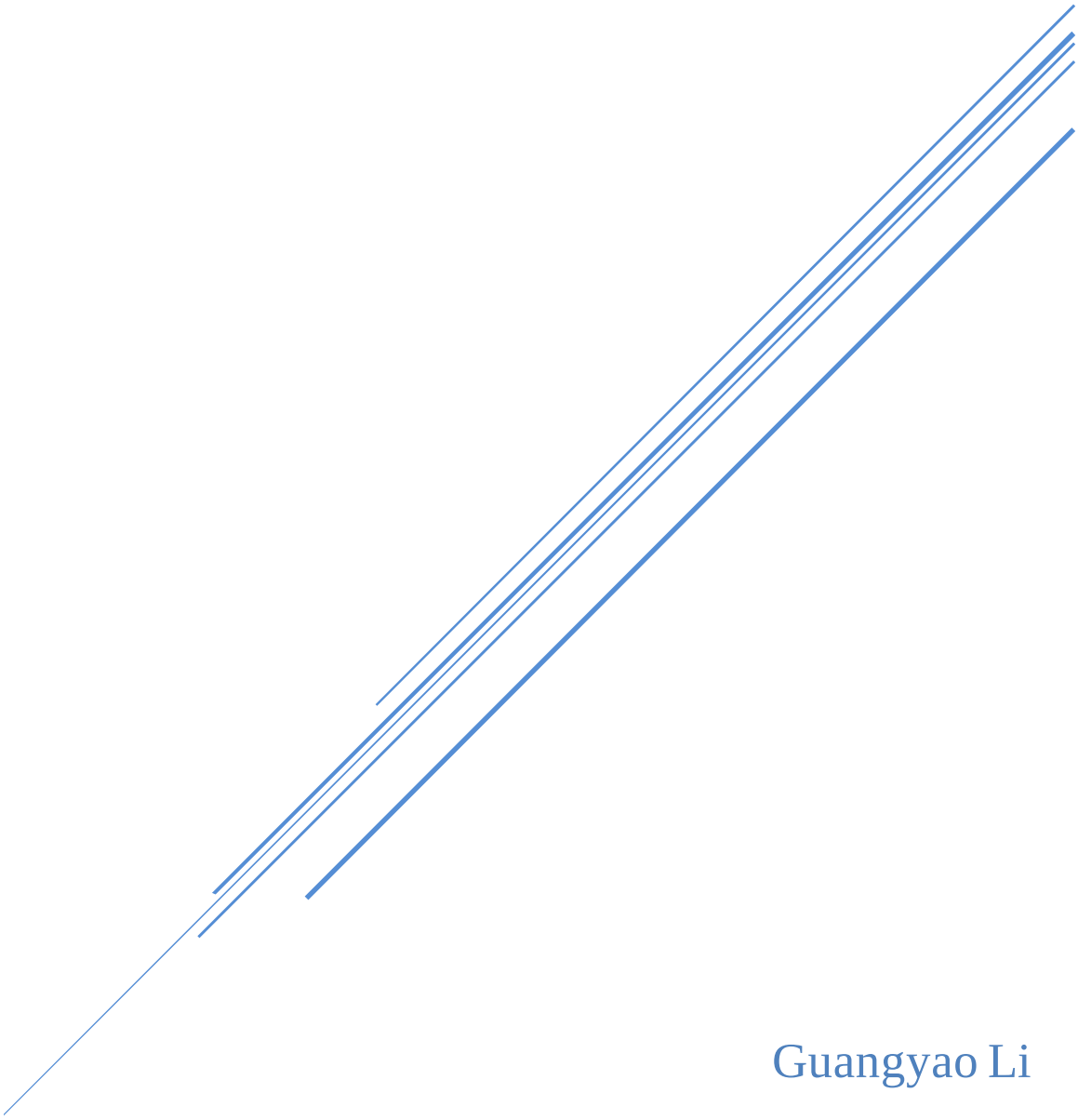
Review BUSA3020 Group Project



Guangyao Li

45496919

Introduction

In order to reduce risk of customer defaults and also maintain profit for the bank, this group report is to estimate the credibility of loan applicants by using statistical models. As a result, we achieved an f-beta score of 0.91 for out-of-sample predictions on Kaggle, which can support our client to predict accurately with a significant degree towards creditworthiness of future loan applicants and assist in better decision making to maximise profits. In addition, my part of the project is to cleanse the dataset by reorganising variables, visualising data, building models (LR & RF) and testing model performance by using Python. I will review this group project from a Business Analytic perspective.

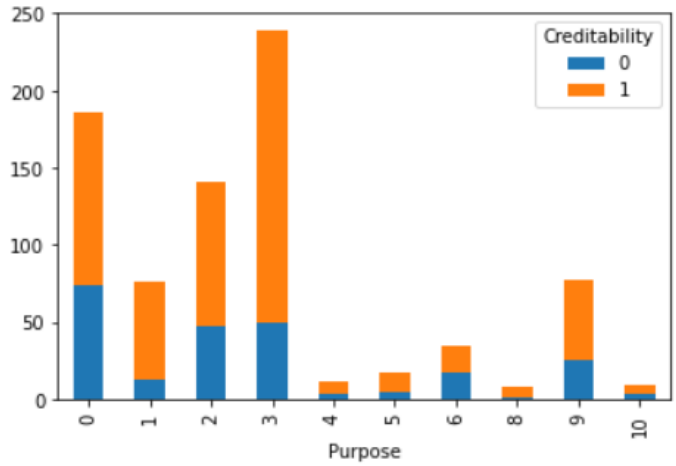
Good characteristics

1. Creativity

When it comes to cleansing the dataset, I found out the variable of '***Purposes'*** owning ***10 sub-variables*** as the following bar chart, which might have disturbed our accuracy of models. Therefore, the tutor told us we could merge them as 'entertainment', 'study', and 'other categorical variables'...

```
pd.crosstab(df['Purpose'], df['Creditability']).plot.bar(stacked=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x139b190ec70>
```

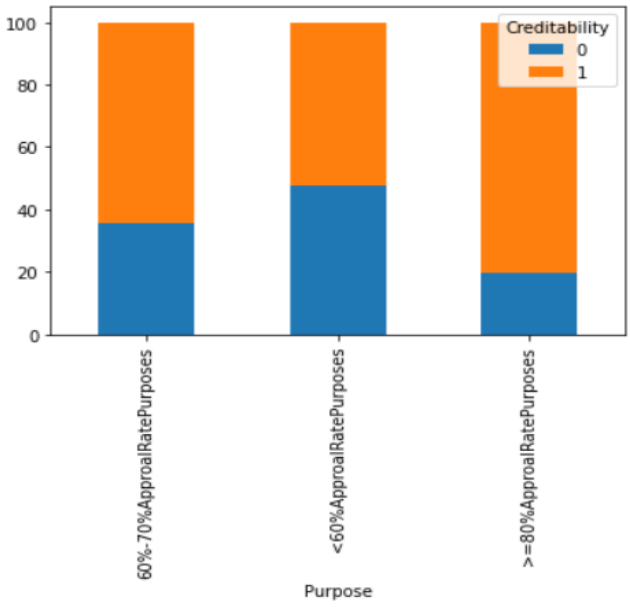


However, I checked out the approval ratio of each purpose through sub-variables ***groupby***, and decided to merge purposes **based on their ratio**:

Purpose	mean	sum
0	0.602151	112
1	0.828947	63
2	0.666667	94
3	0.790795	189
4	0.666667	8
5	0.705882	12
6	0.514286	18
8	0.875000	7
9	0.675325	52
10	0.555556	5

Then, I applied the ***lambda*** function for merging into three more general variables, which likes the following:

```
0:'60%-70%ApproalRatePurposes',
2:'60%-70%ApproalRatePurposes',
4:'60%-70%ApproalRatePurposes',
5:'60%-70%ApproalRatePurposes',
9:'60%-70%ApproalRatePurposes',
1:'>=80%ApproalRatePurposes',
3:'>=80%ApproalRatePurposes',
8:'>=80%ApproalRatePurposes',
6:'<60%ApproalRatePurposes',
10:'<60%ApproalRatePurposes'
```



Which is much more reliable to be applied afterwards and more obvious with the ratio of approval than the previous one, which has 10 variables:

Before:

Purpose	mean	sum
0	0.602151	112
1	0.828947	63
2	0.666667	94
3	0.790795	189
4	0.666667	8
5	0.705882	12
6	0.514286	18
8	0.875000	7
9	0.675325	52
10	0.555556	5

After:

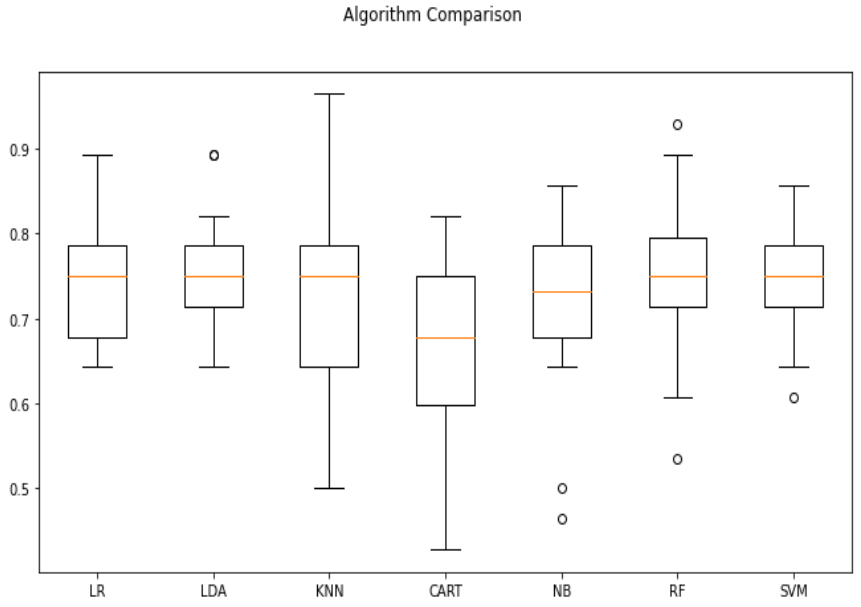
Purpose	mean	sum
60%-70%ApproalRatePurposes	0.642032	278
<60%ApproalRatePurposes	0.522727	23
>=80%ApproalRatePurposes	0.801858	259

To sum up, creativity is based on my confidence in numerical data, which could be turned into better statistical results. Also, I learnt how to use ***lambda*** and ***crosstab*** in pandas and what scenarios should be applied with these functions, which can definitely improve my coding and statistical skills.

2. Evolution of Comparison

When it comes to evaluating each model, I have known how to display them in a effective way, which is ***boxplot algorithm comparison:***

```
LR: 0.748214 (0.068209)
LDA: 0.758929 (0.059628)
KNN: 0.732143 (0.106246)
CART: 0.669643 (0.101566)
NB: 0.717857 (0.093746)
RF: 0.748214 (0.091386)
SVM: 0.746429 (0.061756)
```

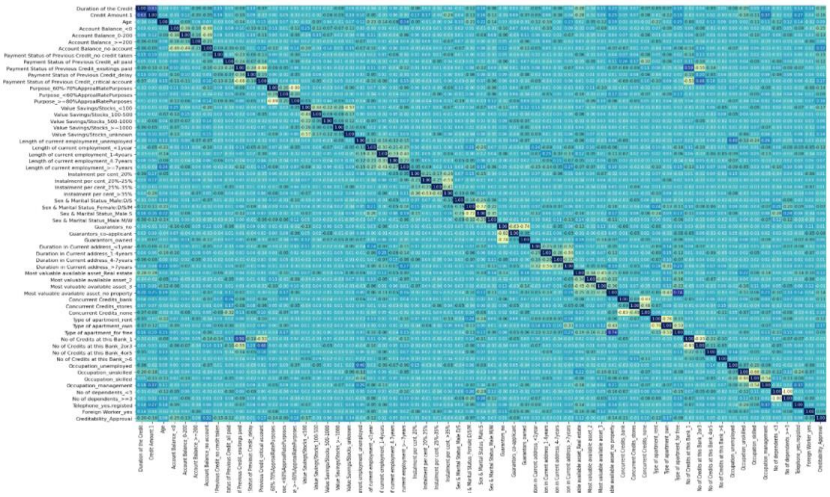


Which is why, I suggested LR and RF as our chosen models and it turned out that their performance was decent.

Room for improvement

1. PCA using

The total variables showed in the notebook as follows:

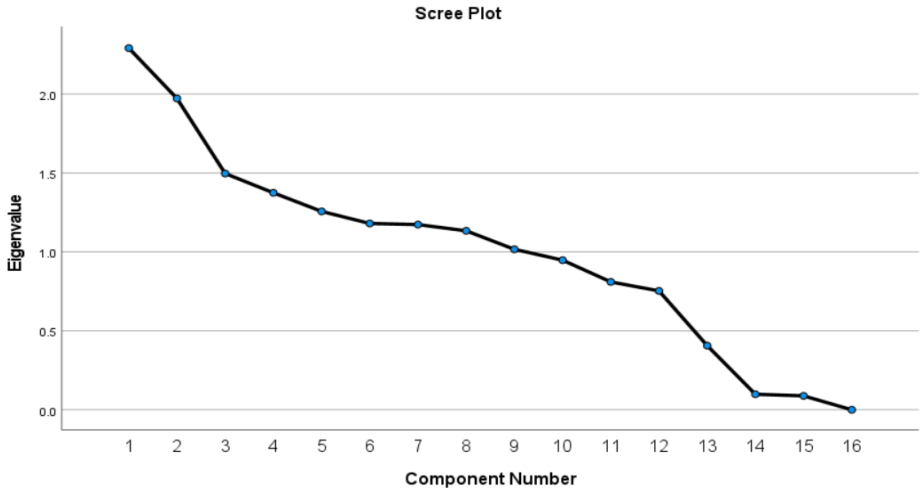


It consists of many unnecessary variables. In order to reduce these variables, the method I applied was to feature importance measures. The final version of variables shows as follows:

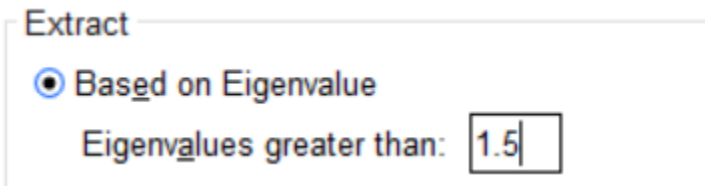
Certainly, the measures can provide strong evidence of which one might be ignored. However, I spent a hefty amount of time on it, deleting a few variables and testing them over and over. In the end, I found out there are still several variables with very low scores, which was not effective and productive to some degree.

As such, after walking through more of PCA, I realised that might help me out by setting Principal Components. Then, I tried to use SPSS to figure it out according to PCA.

Unfortunately, the scree plot shows it has nine PCs that are larger than 1 in terms of Eigenvalues.



If we select two PCs, which means we have to pick Eigenvalues greater than 1.5.



And it can ONLY extract 40% cumulative of variables, so we have to forgo PCA disappointedly. However, identifying the reasons for its failure was what I was looking for.

Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.291	14.321	14.321	2.291	14.321	14.321
2	1.972	12.326	26.646	1.972	12.326	26.646

2. Teamwork

I am not satisfied with my teamwork experience in BUSA3020, especially for the communication with team members. Since everyone just did their work, leading us to lack many opportunities of sharing specific knowledge with each other. Which is why, I expect that all of the group members could be involved in the dataset understanding and designing part, so that we could keep everyone on the same page and gain information and knowledge from others.

Conclusion

When it comes to project completion, there is no doubt that we managed to achieve the goal of the project. When it comes to the process of a project, I gained a few useful techniques and functions which are supportive towards the next project. Also, there is room for me to improve in terms of teamwork skills and leadership. In addition, It is an honour for me to give feedback on my work.

The personal portfolio link is:

https://github.com/GuangyaoLi-45496919/BUSA3020-Group-Project-Portfolio.git