BigData&DataAnalytics

Build and Evaluate Models

Mauricio Carvajal

IOT ANALYTICS

Table of contents

1. Introduction

The following project consist on build and evaluate differetns models of machine learning in order to provide insighs to the customer, and help to take business decision base on data.

In this case the principal problem that this company has is the incrsing default rates that they are having, and this can cause that the bank lost money.

The two main questions that this report will handle are:

- How do you ensure that customers can/will pay their loans?
- Can we approve customers with high certainty?

It’s important to mention that for this project we use 3 models that are familiar, and there is try to start using neural networks algortim, in which this can be dicussed in the following document.

2. Framework that has been used

In the following project, we used python and some important libraries like “sklearn”, “pandas”, “seaborn and matplot” to go to the framework of data science. All the information and comments are in the Jupyter Notebook that has been attaced, but as summary, this the steps that has been follow:

- Cleaning and Preprocesing the data
- EDA (Exploratory Data Analysis)
- Feature Engineering
- Model Development
- Model Tunning
- Model Evaluation
- Model Prediction
- Feature Importance

3 Results

In this section is present the results of the process of build and evaluate the model, in which the most important insights are listed.

3.1 Model build Insights

As in the previous sction the has listed all the steps that has been follow to build and evaluate the model, but here are the insighta of follow up the stepts:

- Cleaning the data is critical, in order to ajust the correctness of the model, for example, chaning numerical values into categorical, process the missing values, removing features that are not relevant.
- In the case of preprocessing the data, this helps to improve the performance of the modelts, in this project we used to scalated the data, grouping or binning in order to understand better the behabiour, for example the age transforming to yougth, yought adults, adutls seniours.
- EDA (Exploratory data analysis), this taks is also critical to understand the data, and the differentet feature and start to having idea the level of importance. But in this project “Gender”

is shown a tendency that impact the default, but when is evaluate trhu the model is not, so it's important to have in mind that this is a initial stage.

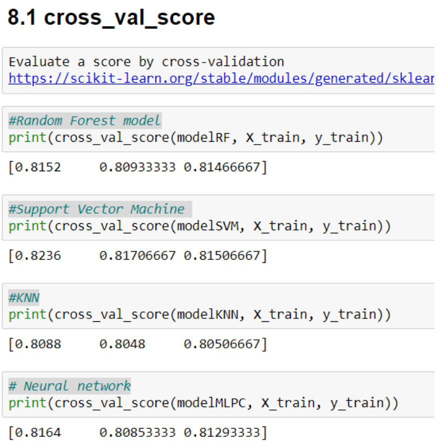
- Feature enginnering is vital, in this project we used PCA that allows to summarize and to visualize the information in a data set containing individuals/observations described by multiple inter-correlated quantitative variables

3.2 Models Restuls

In this Report all the technical information is placed in the "Jupyter Notebook", but something important is to present the results of the models that we used.

As shown in the picture all present similar behavior, but "RF" present better score, so for this project this is the one that is choosen. But the relevant information is the neural network model that not shown better results, and some insights about is:

- We need to do more turn in
- Those models perfrom better with large data sets



8.1 cross_val_score

Evaluate a score by cross-validation
https://scikit-learn.org/stable/modules/generated/sklear

```
#Random Forest model
print(cross_val_score(modelRF, X_train, y_train))
```

```
[0.8152     0.80933333 0.81466667]
```

```
#Support Vector Machine
print(cross_val_score(modelSVM, X_train, y_train))
```

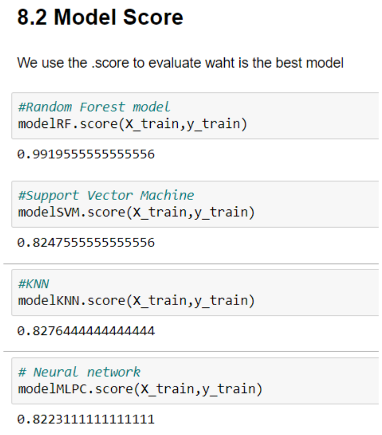
```
[0.8236     0.81706667 0.81506667]
```

```
#KNN
print(cross_val_score(modelKNN, X_train, y_train))
```

```
[0.8088     0.8048     0.80506667]
```

```
# Neural network
print(cross_val_score(modelMLPC, X_train, y_train))
```

```
[0.8164     0.80853333 0.81293333]
```



8.2 Model Score

We use the .score to evaluate waht is the best model

```
#Random Forest model
modelRF.score(X_train,y_train)
```

```
0.9919555555555556
```

```
#Support Vector Machine
modelSVM.score(X_train,y_train)
```

```
0.8247555555555556
```

```
#KNN
modelKNN.score(X_train,y_train)
```

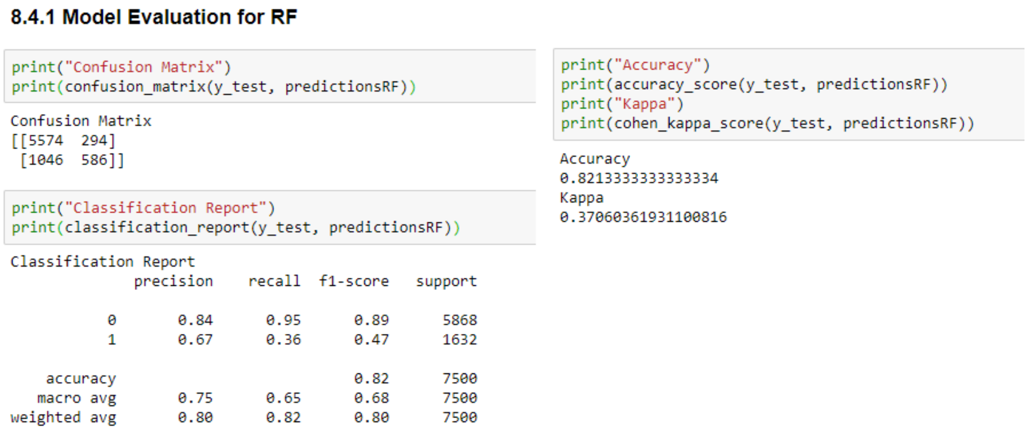
```
0.8276444444444444
```

```
# Neural network
modelMLPC.score(X_train,y_train)
```

```
0.8223111111111111
```

3.3 Model Choosen

In this section, it's discussed the model that has been chosed for this project. In this case the model that present the best performance is the RF model.



8.4.1 Model Evaluation for RF

```
print("Confusion Matrix")
print(confusion_matrix(y_test, predictionsRF))
```

```
Confusion Matrix
[[5574  294]
 [1046  586]]
```

```
print("Classification Report")
print(classification_report(y_test, predictionsRF))
```

```
Classification Report
              precision    recall  f1-score   support

           0       0.84      0.95      0.89      5868
           1       0.67      0.36      0.47      1632

    accuracy                           0.82      7500
   macro avg       0.75      0.65      0.68      7500
weighted avg       0.80      0.82      0.80      7500
```

```
print("Accuracy")
print(accuracy_score(y_test, predictionsRF))
print("Kappa")
print(cohen_kappa_score(y_test, predictionsRF))
```

```
Accuracy
0.8213333333333334
Kappa
0.37060361931100816
```

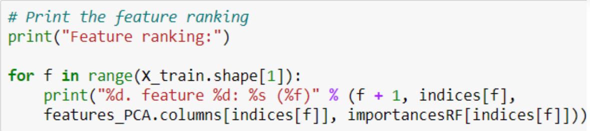
But evenf or this model it’s shown that the kappa value is not representative value to evaluate a good model, the same behavior is shown in the confusion matrix and in the classification report. Even the accuracy of this model is good value, but as overall summary this model can be improved with more tunning.

4.4 Feature importance

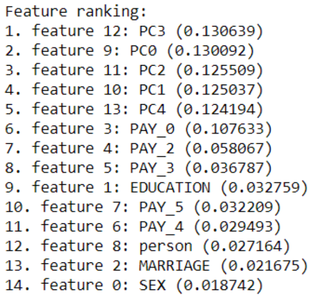
After build an evaluates the 4 models that has been develop in the project, we analasys the data of “RF” which is the one that has been choosen. In this case we noticed that the Top five features that are more important to define is the next month pay will be default or not, are related with the amount of bill statement.

So this amount that we used in this project has been used PCA in order to improve the efficiency of the model, but this represent a combination of present bill and the previous amount that has been payment.

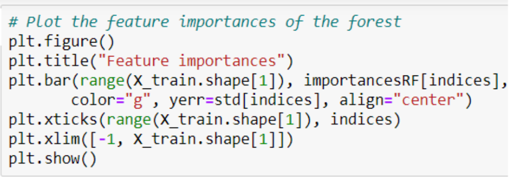
From this model we see that factors like “Gender” or “marriage status” are not relevant to predit if the person will defulat the next month payment.



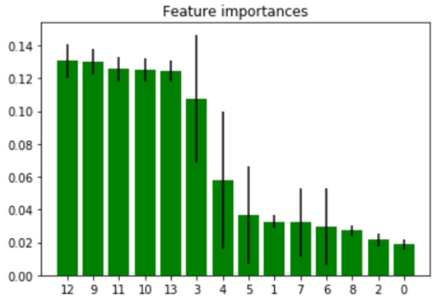
```
# Print the feature ranking
print("Feature ranking:")

for f in range(X_train.shape[1]):
    print("%d. feature %d: %s (%f)" % (f + 1, indices[f],
    features_PCA.columns[indices[f]], importancesRF[indices[f]]))
```



```
Feature ranking:
1. feature 12: PC3 (0.130639)
2. feature 9: PC0 (0.130092)
3. feature 11: PC2 (0.125509)
4. feature 10: PC1 (0.125037)
5. feature 13: PC4 (0.124194)
6. feature 3: PAY_0 (0.107633)
7. feature 4: PAY_2 (0.058067)
8. feature 5: PAY_3 (0.036787)
9. feature 1: EDUCATION (0.032759)
10. feature 7: PAY_5 (0.032209)
11. feature 6: PAY_4 (0.029493)
12. feature 8: person (0.027164)
13. feature 2: MARRIAGE (0.021675)
14. feature 0: SEX (0.018742)
```



```
# Plot the feature importances of the forest
plt.figure()
plt.title("Feature importances")
plt.bar(range(X_train.shape[1]), importancesRF[indices],
       color="g", yerr=std[indices], align="center")
plt.xticks(range(X_train.shape[1]), indices)
plt.xlim([-1, X_train.shape[1]])
plt.show()
```



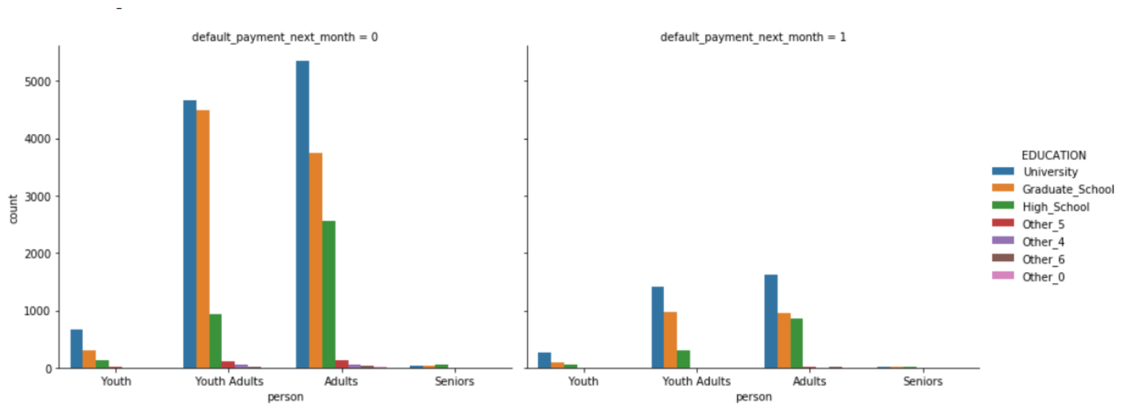
4. Anwers base in data

In this section we discuss the 2 main question that the client is requested, this is base on the data analsys perform, but for further technical details please go to the “Jupyter Notebooks”.

4.1 How do you ensure that customers can/will pay their loans?

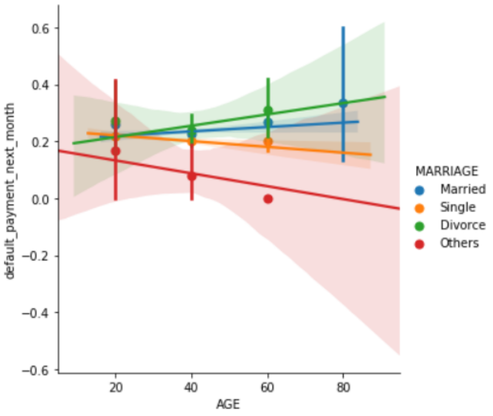
Base on the data that is shown bellow in the feature importance, the top five feacre are related as the amount of bill statement, but at this is previos all the strategy can not be change the customer spending habits.

Another relevant is tha the Education and person feacture that those ca be control by the bank in their target publicity, so this can be usefull that that if you section that has less risk will be very usefull, the next picture showns this discussion.



4.2 Can we approve customers with high certainty?

Base on the data, and the results shown in the below, the model that has been develop give us 82% of accuracy to predict if the customer will be default in the next month payment, but the kappa is not fully support. So the bank can take the risk to do this actions but can be minimize including more % of interest.



5 Recommendations

- The models can be improved by more rounds of tunning, but this can be dangerous in order to overfit the model.
- The neural network model that has been used, not perform better than regular machine learning algos, this probably due to need more tunning and the data set is relative small for this model
- Preprocess the data is Have to do activitiy, as same as feature enginnering to help the models to predict better.
- For improvement this data, also can request to the cliend more information that can be usefull to see is more relationship
- Predict the human behaours is not easy, and the most studies suggest that is difficul find good metrics.
- EDA has very usefull to have graphics and pictures on the data, that is very easy to explain to manameent or high executives of the company.