1. Task Description

Hi,

Now that you have successfully imported, prepared and explored the data you are ready to start exploring some possible tools for your analysis. When you worked in R you used the caret package for many of your machine learning and data mining tasks. Python has a similar library called Sci-Kit Learn that the client has specifically asked us to use because it is likely to be compatible a custom software solution they plan to deploy.

In this task you'll build your models just as you have done previously, but with a different set of tools. As you progress remember the following:

Let the data tell the story – don't make any assumptions. It is often best to build three or more models and compare the results. Make sure you have chosen the correct tools for the type of data you have. I suggest you start this task with a quick orientation on Sci-Kit Learn to become familiar with the benefits of using it and how to use it effectively for this project.

Guido is expecting a report in a few days:
1 - Your report should be a one to three-page Word document that includes rules you believe provide insights, any relevant visualizations, and the answers to the company's questions.
2 - It should also include any observations that you've made and any recommendations you might have, supported by evidence uncovered in your analysis.

Here is the list of requirements that your data science process should include for your final report:

1 - Cleaning and Pre-processing
2 - Covariance Estimation
3 - EDA
4 - Feature Engineering (either PCA or RFE) and Dimensionality Reduction
5 - One-Hot Encoding (if needed)
6 - Classification (Build three model and choose the best)
7 - Model Tuning (Tune at least two parameters for each model you build)
8 - Model Evaluation

GR

Guido Rossum Senior Data Scientist Credit One
www.creditonellc.com

2. Task Solution

One data set (default of credit card clients.csv) were provided by Guido Rossum, to conduct feature engineering and model development to predict if customers will default their future loan payments. The best model will be selected (based on highest accuracy, precision, recall and F1-score values) to conduct the predictions. Five models (Logistics Regression, KNN, CART, NB and SVM) will be developed and tuned to enhance the model and get predictions with higher accuracy.

2.1 Models Configuration

Models configuration: Each model contains the following processes and parameters:

- Pre-processing
- Data Partition (70/30)
- Kfold (n_splits= 10 random_state = 7)
- Post Resample
- Predictions

Selected Algorithms:

1- Logistics Regression (LR)
2- K-nearest neighbor (KNN)
3- Classification and Decision Tree (CART)
4- Naïve Bayes (NB)
5- Support Vector Machine (SVM)

2.2 Data preprocessing, modelling and evaluation

The flow chart listed below (Figure 1) shows the preprocessing steps defined to subset the data, build and tune all the algorithms, the model evaluation is used to determine the best performer algorithm.

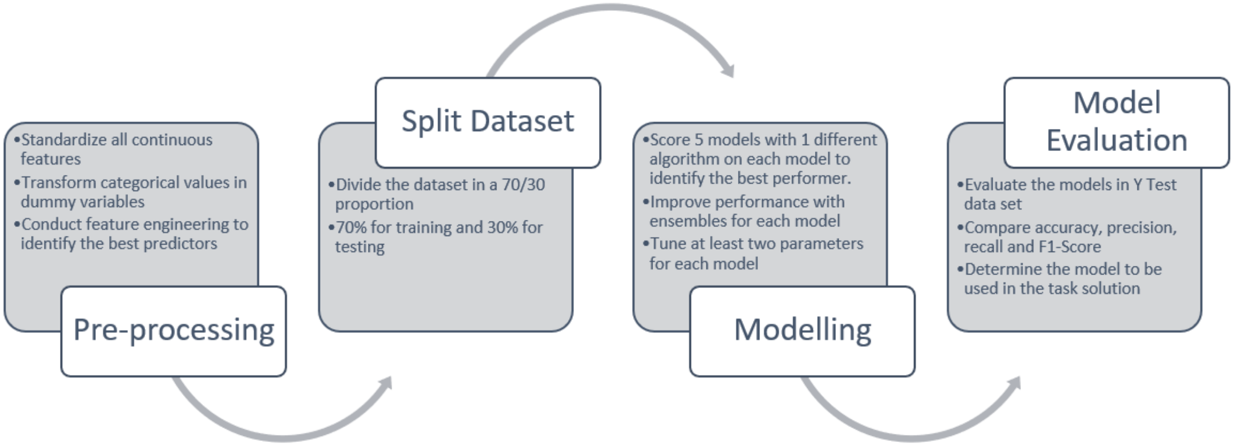
•Standardize all continuous features
•Transform categorical values in dummy variables
•Conduct feature engineering to identify the best predictors
Pre-processing
Split Dataset
•Divide the dataset in a 70/30 proportion
•70% for training and 30% for testing
•Score 5 models with 1 different algorithm on each model to identify the best performer.
•Improve performance with ensembles for each model
•Tune at least two parameters for each model
Modelling
Model Evaluation
•Evaluate the models in Y Test data set
•Compare accuracy, precision, recall and F1-Score
•Determine the model to be used in the task solution

Figure 1 – Workflow for Credit One Dataset

3. Results

3.1 Model Results

Based on the steps proposed in section 2.2, the modelling step provided the following outcomes. The LR, KNN, CART, NB and SVM models were fitted and scored under the specified conditions, the out of the box model with the best performance is the Linear Regression (see Table 1 and Figure 2).

Model Name	Accuracy
Logistic Regression (LR)	**0.8200**
K-nearest neighbor (KNN)	0.7959
Classification & Decision Tree (CART)	0.8012
Naïve Bayes (NB)	0.7944
Support Vector Machine (SVM)	0.8182

Table 1 – Accuracy values for model scoring

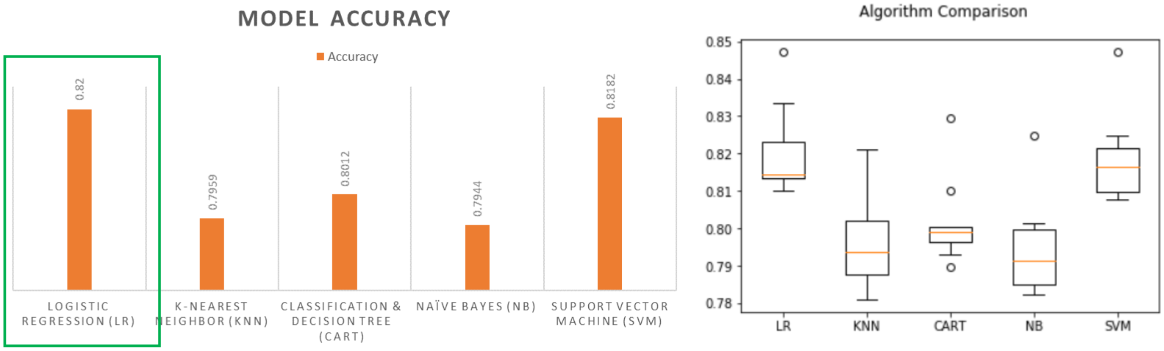

Figure 2 – Accuracy values for out of the box models

The first round of performance improvements is based on ensemble methods, four classifier methods were selected: stochastic gradient boosting, random forest, extra trees and bagged decision. The best performer is **Stochastic Gradient Boosting** with an Accuracy value of 0.8205.

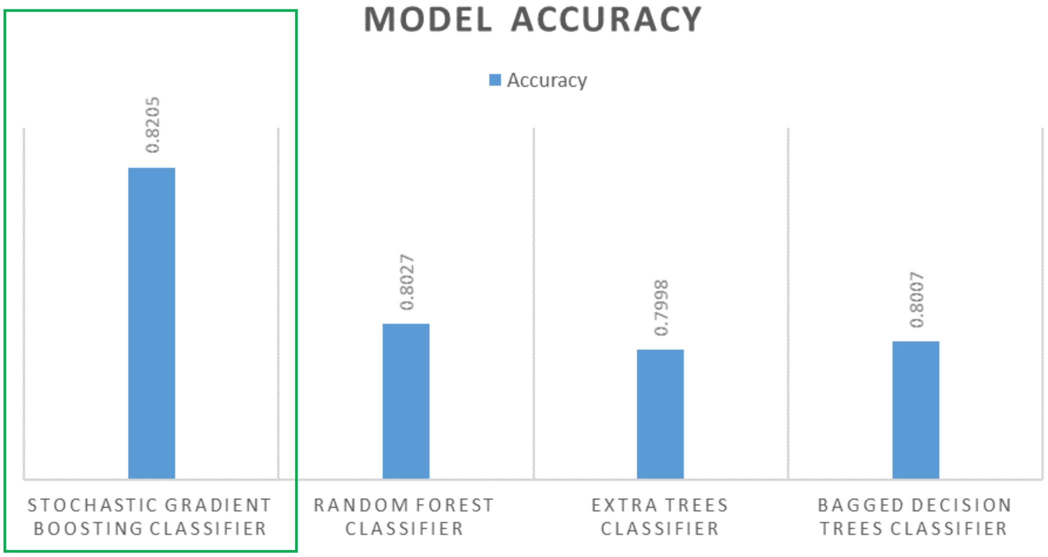

Figure 3 – Accuracy values for ensemble methods

The second round of performance improvements was based on model tuning, at least two tuning parameters were modified for the top 5 models (Stochastic Gradient Boosting, Logistic Regression, Support Vector Machine, Decision Tree and KNN). See figure 4.

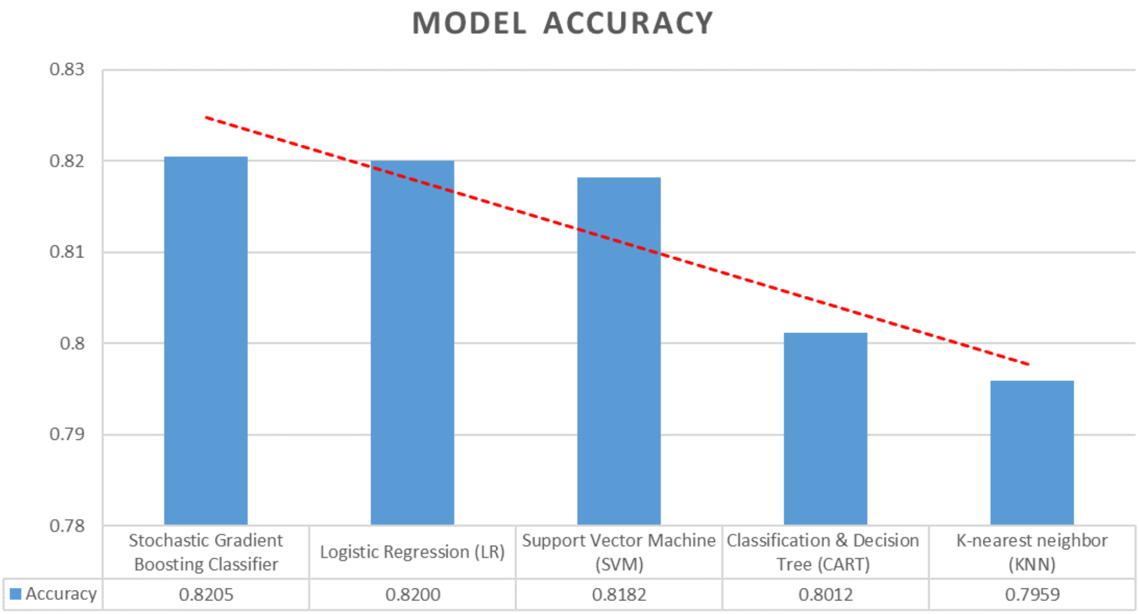

Figure 4 – Accuracy values for selected models

The tuned parameters of the models show some degree of improvement in 4 out of 5 models. As is show in figure 5, in this process the best performer was the Logistic Regression.

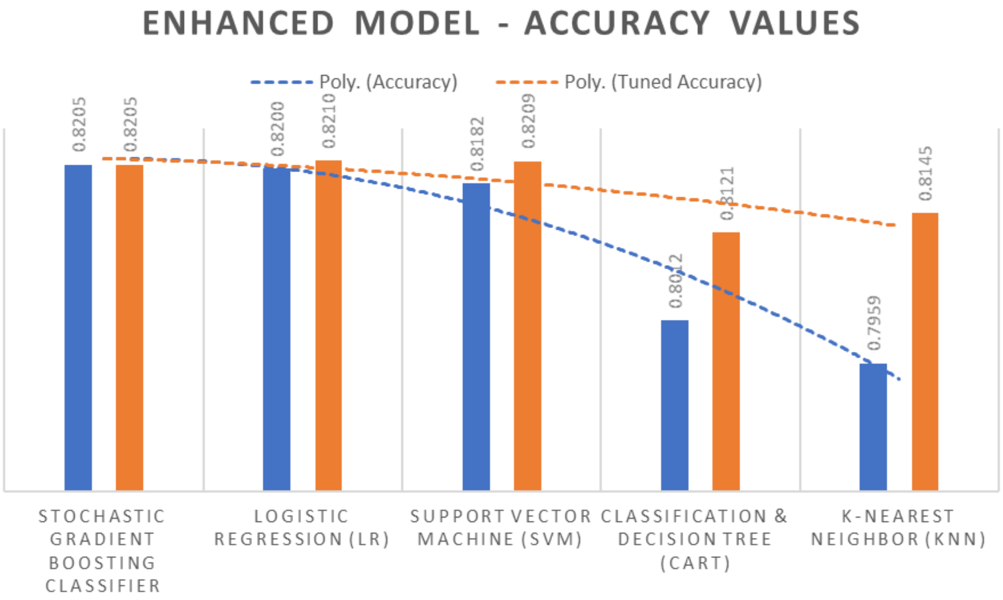

Figure 5 – Accuracy values for enhanced models

All models were used to conduct predictions (model details can be reviewed in the jupyter notebook). The results of the predictions are listed below in Table 2.

Model Name	Accuracy	Precision	Recall	F1-Score
Stochastic Gradient Boosting Classifier	0.82011	0.80	0.82	0.80
Logistic Regression (LR)	0.81866	0.80	0.82	0.80
Support Vector Machine (SVM)	0.81888	0.80	0.82	0.80
Classification & Decision Tree (CART)	0.79688	0.77	0.80	0.78
K-nearest neighbor (KNN)	0.81633	0.80	0.82	0.79

Table 2 – Accuracy values for model scoring

The results obtained from the predictions, shows that Stochastic Gradient Boosting mode has the highest Accuracy (0.82011) and F1-Score (0.80).

4. Results Discussion

Stochastic Gradient Boosting model is the best model we suggest Credit One Company to use, because this model has the highest accuracy and F1-score. The accuracy of this model is 0.82011, and the F1-score of this model is 0.80, which is good because is closer to 1.

Accuracy is used when the True Positives and True negatives are more important while F1-score is used when the False Negatives and False Positives are crucial.

In our case both metrics are important. However, F1 score gives a better measure of the incorrectly classified cases than the Accuracy Metric.

5. Recommendations

If the accuracy must be improved, additional data can be collected to update the model.

6. Python scripts

Provided in Jupyter Notebook and Github link.