**Skill-based Assignment Report**

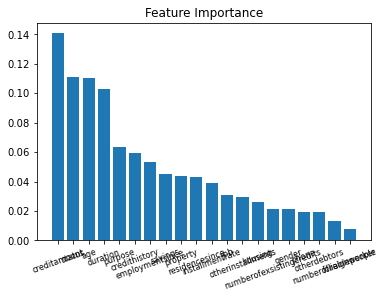
**Q1) Most important determinants of credit risk application: -**

**a) Present convincing arguments based on insights from the dataset**

Based on observation of the Customer dataset alone, I believe features like Credit History, Age, Savings, Employment Since could be important for predicting the target variable – Credit Worthy column in this case.

I have shortlisted the independent/ predictor variables (Feature Selection) based on *Feature Importance* technique, which provided an insight into the dataset and helped me with choosing the relevant features for training the model.

Upon feeding all input/ predictor variables, the following features were ranked better - *Credit amount, Status, Age, Duration, Purpose, Credit history, Employment since*



**(b) What are the limitations of your argument?**

Since we don’t have much data sample (988 rows) for training models, potentially the trained models can’t be deployed into production or in a live application. Techniques like Cross-validation might have to considered as well rather than using the conventional Test Train Split method, as we are working on a limited data sample.

**Q2) The management team wants to use your analysis to propose a system. Can you give an estimate of how accurate your algorithm is likely to be?**

Since we are attempting to predict a categorical result (i.e., if an applicant is credit worthy or not), we will be implementing a supervised classification algorithm.

In this case, I have decided to use the following algorithms and do a comparison between these techniques

* *k-nearest Neighbours*

*As it’s one of the simplest, and faster classification algorithms that requires less training.*

* *Linear Support Vector Machine*

*Though it’s usually used in high dimension problem scenarios, I wanted to analyse how well it could fare in comparison to other techniques. SVM models also tend to perform better on sparse data as well.*

* *Random Decision Forest*

*Decision trees might be the default approach for a classification problem like this. Implementing random decision forest, as it’s an ensemble of decision trees and it’s likely to be less prone to overfitting and more accurate.*

The samples were found to be unique enough and there weren’t any NULL values in the records that had to be cleansed. However, the target variable (the credit worthy column) had to be transformed into an integer from textual form.

Based on my experiments, Random Forest had the better prediction score (~77% at times) amongst the three classification algorithms.

**Q3: What according to you are some of the benefits and risks of adopting such analytics in assessing credit applications.**

Introducing machine learning into such applications could reduce human intervention and also, as it could improve and evolve after deploying, predictive analytics are well suited for such applications. The benefits would in-term be less human made error.

But at the same time, such applications would require large datasets in order to properly train a model and tasks like understanding business requirements, shortlisting parameters, hyper parameter tuning, pre-processing and transforming data, training model and testing could all be bit time consuming.