# Data Academy project feedback

# Learner: Thomas Langlois

# Marker: Robert Jirschik

# Module: Classification

# Indicative grade: Pass

# Overall comments:

Good project, even with the negative results which are still results and delivered some insights as to the importance of your features in relation to the KYC rating.
This project could be taken from decent to fantastic by obtaining more relevant data to predict the KYC rating, such as publicly available data from the World Bank. The visualisations and presentation in general were great! It would also be great to explore further Classification algorithms beyond the DecisionTreeClassifier.

# Data pipeline

The table below contains specific feedback on each section of the data science pipeline.

| Stages | Feedback |
|---|---|
| Creating a data question<br>*Collect, manipulate, and collate data from a range of sources to solve specific problems* | **Strengths**<br>+ Clearly identifies an appropriate data question and states testable hypotheses and initial expectations<br>+ Sources relevant data and explains its application to the organisation<br>**Improvements**<br>- Data sourcing could have been improved: Use publicly available data about countries considered |
| Exploration<br>*Use statistical tools and visualisations to explore & summarise data* | **Strengths**<br>+ Examines and discusses core features of the data<br>+ Uses a range of well-chosen statistical methods to analyse data<br>+ Explores in detail the relevance of identified relationships and uses insights to guide the project<br>**Improvements**<br>- Could have used a range of visualisation methods and interprets the results with detail and clarity, e.g. bar/pie charts to visualise country or client type |
| Preparation<br>*Deal effectively with data issues and inconsistencies* | **Strengths**<br>+ Well done to change the data type of the label from numerical to categorical<br>**Improvements**<br>- Could have explained that DecisionTreeClassifier only works well for categorical data (Especially since you clearly know this) |
| Analysis | **Strengths** |

| | |
|---|---|
| *Use a range of analytical techniques to model and understand the data* | + Selected an analytical technique to model the data<br>+ Successfully employed an analytical technique<br>+ Caught the problem of imbalanced data<br>+ Judged statistical metrics of confusion matrix / accuracy / precision / recall very well<br>**Improvements**<br>- Obtain data with more features. Could have for example been combined with publicly available data such as OECD / World Bank / some corruption index to get a better prediction<br>- Could have used other Classification algorithms (e.g. Naive Bayes)<br>- Could have experimented and played around with the decision tree parameters |
| Interpretation<br>*Draw clear conclusions and insights from data* | **Strengths**<br>+ Draws detailed and well-considered conclusions from the model<br>+ Explores the limitations of the model and its applicability beyond the data<br>+ Fully considers the data question and the hypotheses<br>**Improvements**<br>- Could have given suggestions for further research |
| Presentation<br>*Clearly communicate ideas, insights and processes* | **Strengths**<br>+ The project is well-structured & presented consistently<br>+ The project has a clear narrative from beginning to end and is comprehensible by a reader without specialist knowledge<br>+ The text is well articulated and engaging<br>+ The visualisations are fantastic<br>**Improvements**<br>- Could have visualised the decision tree with graphviz |