Machine Learning for Data Analytics

Assignment 2

This design document explains the problem undertaken, the data to be worked with and features associated with it, any problems found within the data, the decisions taken to tackle the problem and the steps leading to the final solution itself. The classification goal is to predict if the client will subscribe to a term deposit after they have been contacted by phone.

The data is related with direct marketing campaign of a banking institution based on phone calls. Often more than one contact to the same client was required, in order to access if the product (bank term deposit) would be subscribed or not.

The attributes associated with the data are as follows: id, age, job, marital, education, default, balance, housing, loan, contact, day, month, campaign, pdays, previous, poutcome and subscribed. A description of the features are as follows:

* ID – Unique identifier of the row of data
* Age – numeric value describing the age of the person in retrospect to the row queried. Best described as a continuous feature of the user. The median age for the category is 37, the mean is 40 (rounded to nearest) and the mode is 32 years old. After further looking into the data, the general description of the outliers for the dataset are in the elder category (60+) and people in the Young Category (22 or younger).
* Job - type of job associated with the user. Best descripted as a categorical feature, as the jobs are categorised in 12 different levels ('JobCat1' to 'JobCat11', and 'unknown'). Since there are limited number of categories the bank has restricted to in description, the data can be interpreted as continuous features by configuring them.
* Marital - marital status of the user. Best described as a categorical feature as the descriptive features matches category standards. The choices for this application are "married", "divorced" and "single". In the case of "divorced", it could be noted that the interpretation can be divorced or widowed. When looking at the data, there is a gap between the three different choices, where between the married and single contacted, there is a difference of 812 members, where as the difference between single and divorced is 500 members. This would indicate clearly married people dominate the set.
* Education – The highest educational level completed by the individual. This is best described as a categorical feature. The choices are "unknown", "secondary", "primary" and "tertiary". The education status in the dataset contains disparity between the members, where completion of secondary is 595 more than tertiary education and is the dominant value of the feature. Tertiary itself is 397 more then primary education. A definite outlier is the unknown feature, as they constitute only 102 members out of 2700.
* Default – Description of the status of the credit (ie is the credit in default?). Since this a binary "yes" or "no", the feature could fit into either continuous or categorical description, as the binary factors could be changed for 1as yes and 0 as no. However, this data contains only 45 yeses’, indicating majority of people do not have credit in default.
* Balance – The average yearly balance, in euros, the person has saved up. Since we are dealing with numeric values, these are best described as continuous features. These features are the only part of the application with the highest number of outliers within the dataset. Majority of people have only minimal amount stored, which goes no higher than early to mid-section of the triple numeric digits. It is also indicated people have money stored just so they have an account with the bank itself. While the mean of the data is 1396, the mode is only 205 and the median is 442. This indicates there are a few minority people having accounts that stored between 5 to 50 times their counterparts, which alters the data significantly.
* Housing – Does the client have a housing loan currently? Since this is only a binary yes or no, the feature could be described as a continuous or categorically feature. There is an expected balance between the two choices, with yes only holding approximately 8% more places.
* Loan – Does the person have a personal loan needed to pay off? Since this is only a binary yes or no, the feature could be described as a continuous or categorically feature. Since people whom have loans sums up to be 400 out of 2700, this would have a significant role in the predications area about the dataset.
* Contact – The communication type used to engage with the client. The choices are "unknown", "telephone" and "cellular". This feature is a categorical example. It is known that majority of people have a cellular phone, due to the dominant 64% of the set combined with everyday knowledge of the handiness of the set.
* Day – The day of a given month when the client was last contacted. This is a continuous feature with a dependency on the monthly feature.
* Month – The last month in which the client was contacted on. The choices are January to December. While it is convenient to use this feature as a categorical one, the limited range means this could be used to describe the data as a continuous method. While the disparity between the most likely day to be called on and the least could be argued about, the steps of the depreciation of the values between the highest to the lowest indicates the outliers have minimal impact of the dataset. The outliers on the lower end of the spectrum are Jan, Oct, Sept, March and Dec, in which is known for the specific universal events all around, such as Christmas. The only month impact the dataset from a higher perspective is May.
* Campaign – The number of times the client was contacted during the campaign. Since this is a numeric value, it is best described as a continuous-features. This number would include the last contact made to them. Data indicates very few get bombarded with multiple contacts, which greatly impacts the possibility of being subscribed.
* Pdays – The figure containing the number of days passed by after the row’s client was last contacted from a previous campaign. If the number amounts to -1, the client was not previously contacted. Since this is dealing with figures, it is regarded as a continuous feature. Pdays is short for Previous days since contacted. Since looking into the dataset, it was found that 83% of the people are new to the campaign, as they have never been contacted beforehand.
* Previous – The amount of times the client in the given row was contacted before this contacted was performed. Since this is a numeric value, this is regarded as a continuous feature and aligns with the data found through pdays.
* Poutcome – The outcome of the previous marketing campaign performed on the person. Since majority of the people were not contacted before, the success rate could not be calculated from the dataset.
* Subscribed – The actual variable where the predication of the bank being subscribed was stored.

All the mentioned features, except the subscribed feature and the ID as it’s a unique identifier for each element, are input features used to calculate the possibility of the client to be subscribed, which is the output feature. After checking out the data to ensure there was no missing values from the inputted, such as getting the count of null values, the only issue was the possibility of the dataset being impacted by the outliers in the dataset, which would be minimal overall.

The classification model chosen for the dataset was Gaussian Naive Bayes Classifier from the Scikit-learn python library. Gaussian Naive Bayes Classifier is software that implements the Gaussian Naive Bayes algorithm on continuous features. As a subset of the Naïve Bayes Classifier, Gaussian use methods of supervised-learning style which induces each feature has conditional independence between each other. This, combined with the decoupling of the software and proven workings on real life projects, are the main reasons for the priority of its selection. The absolute final reasoning for the decision to adapt the classifier to our project was insight into these examples showed high level of similar relations between the researched project and our own. After running a few tests through the example provided, it provided satisfactory results with the workings of the set. Due to time constraints, modelling a classifier from scratch was not an option for the dataset.

Issues with the data was the combination of continuous and categorical features. The model did not directly support applying predication on categorical features. To overcome this, a library that allowed the conversion of pure categorical features to continuous features was imported. This was found through the pre-processing library included in the Scikit-learn python library. This library created unique labels for any unique feature values passed through, effectively converting them to continuous features. Once that happened, the capability of using the model to take any feature needed to train it became readily available.

The testing of the data was started after the model was successfully. It was agreed between the team that all the features, apart from the id and the output feature, would be used to generate the predications against the model. To train the model, data was pulled from the file which already was successfully generated. This was to ensure the predication level was the focus before running the model against the file that needed the generation to be completed. Different amount of the data was trained against, such as 33% and 45%, to ensure the highest possible accuracy was found. In the case of the chosen model, 45% of the model produced the highest accuracy when training the model, at a total of 80%. Any higher or lower would create a depreciated solution in comparison.