Banking – Predicting if Potential Clients will Join the Bank

DSC680 – Project 1: Milestone 1

Clayton Evans

The focus of the project will be to predict if by using various customer characteristics if they will become clients of the bank. This can help the bank better target their marketing efforts and to help to maximize the money spent to attract new customers.

The technical elements of this project will focus on:

- Cleaning and preparing data
- Exploring and visualizing data
- Model Selection
- Improving machine learning model performance

The dataset includes 17 variables related to a direct marketing campaign of a Portuguese banking institution.

The data source I will be using can be found at:
https://archive.ics.uci.edu/ml/datasets/Bank%20Marketing

The variable breakdown is as follows:

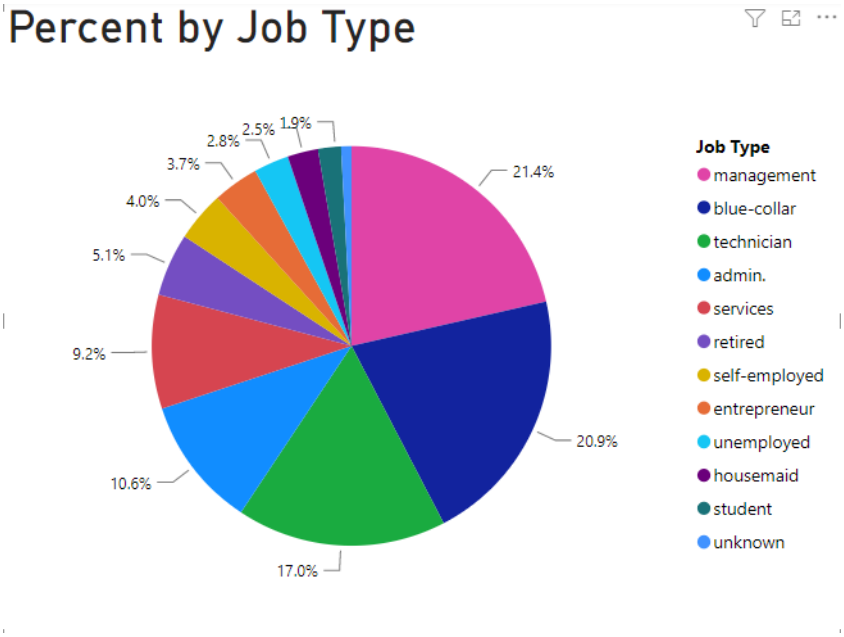
The input variables will be:

1 - age (numeric)
2 - job : type of job (categorical:"admin.","unknown","unemployed","management","housemaid","entrepreneur", "student", "blue-collar","self-employed","retired","technician","services")
3 - marital : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)
4 - education (categorical: "unknown","secondary","primary","tertiary")
5 - default: has credit in default? (binary: "yes","no")
6 - balance: average yearly balance, in euros (numeric)
7 - housing: has housing loan? (binary: "yes","no")
8 - loan: has personal loan? (binary: "yes","no")
related with the last contact of the current campaign:
9 - contact: contact communication type (categorical: "unknown","telephone","cellular")
10 - day: last contact day of the month (numeric)
11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
12 - duration: last contact duration, in seconds (numeric)
other attributes:
13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
15 - previous: number of contacts performed before this campaign and for this client (numeric)
16 - poutcome: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

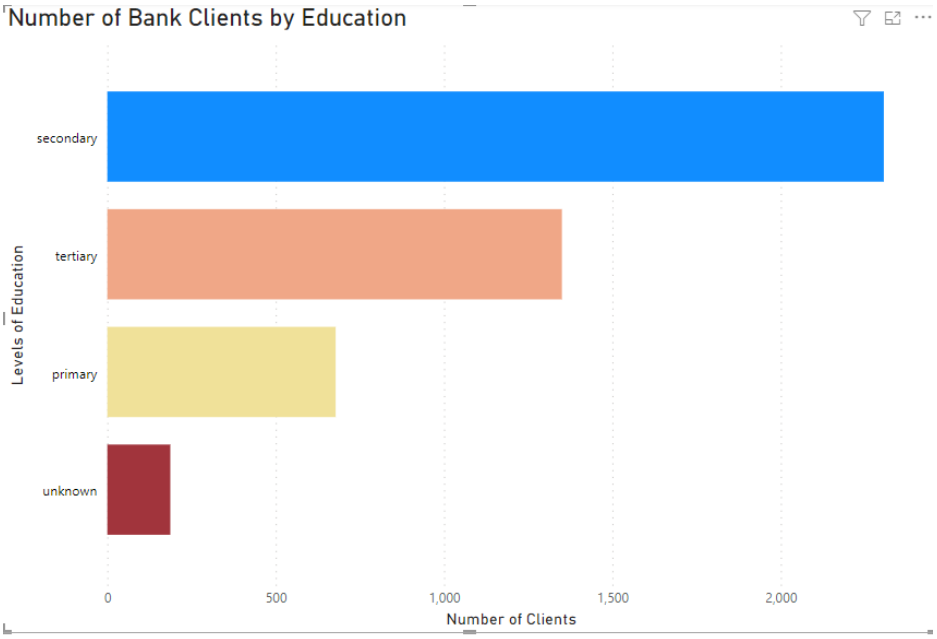
Output variable (desired target):

17 - y - has the client subscribed a term deposit? (binary: "yes","no")

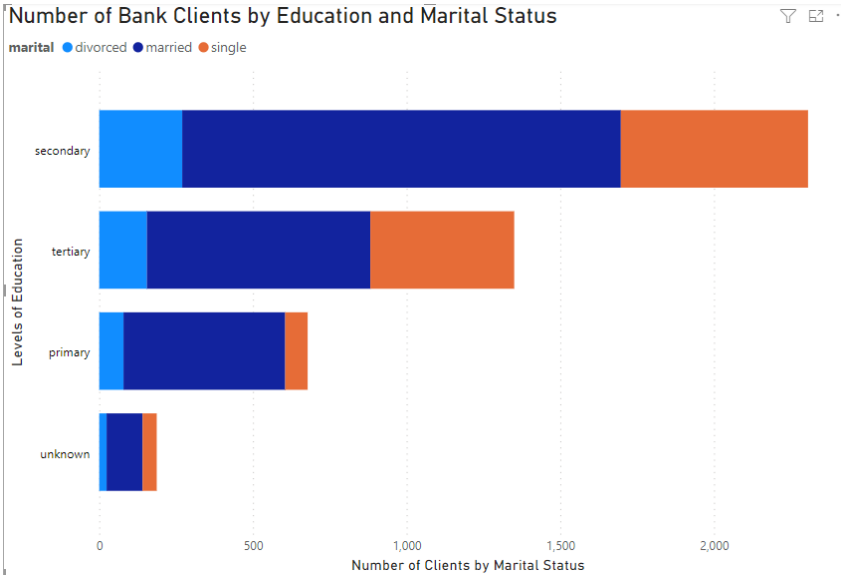
Methods used in this project will be predictive in nature and bet tested for accuracy. The analysis will involve all the variables as they can together be quite predictive. For example seeing the distribution of the type of jobs the potential clients have like below can be helpful:



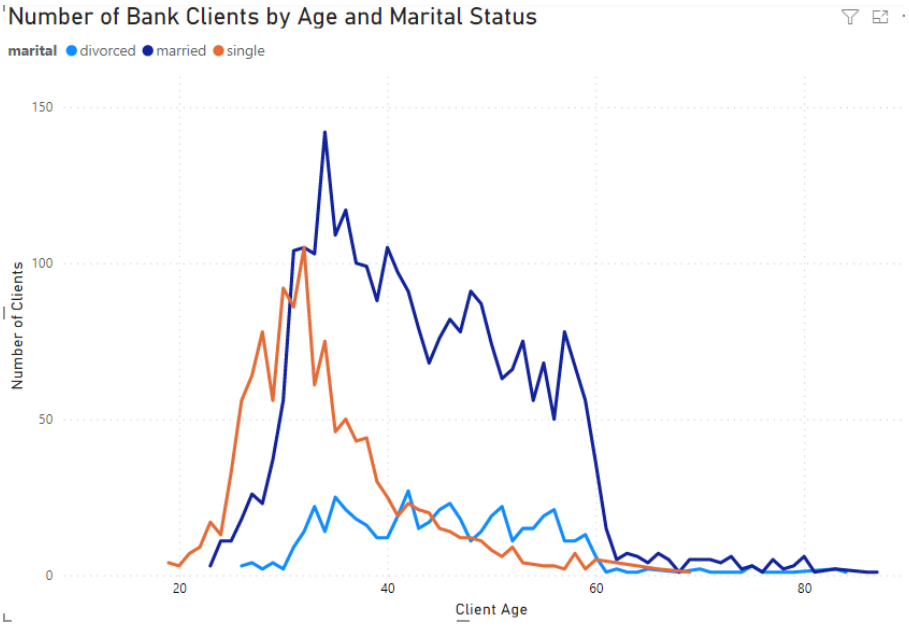
Also, the number of clients by education like below:



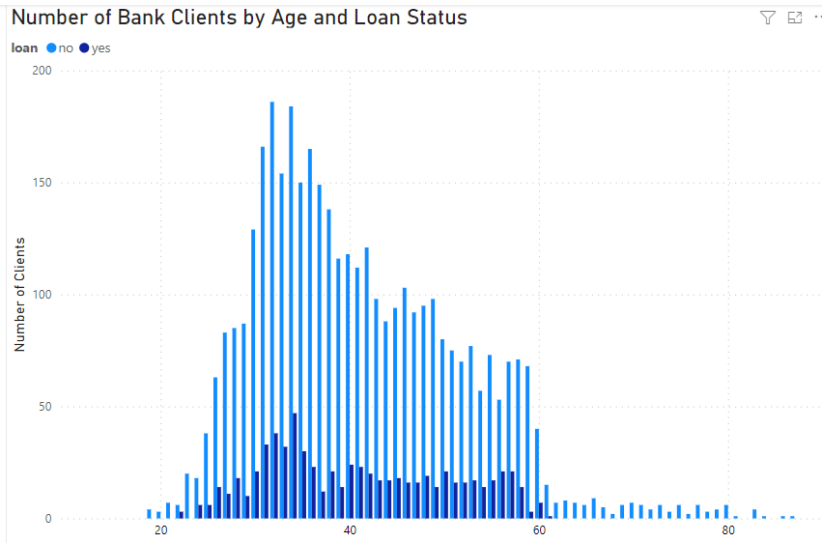
There can be greater impact though by combining variables like the potential client's education and marital status like below:



How visuals are prepared can be of impact too as a line graph of the potential client's age and marital status as below:



Plus combining histograms can be of value to quickly see comparisons like the potential clients age and loan status:



The conclusion though is that even though the above can be quite telling, it can not quite predict if a potential client will join the bank. This is where being able to combine all the variables in a predictive model will be of value.

Assumptions will be that there will be a need for cleansing as well as for standardizing the values. It is also being assumed that by using all the variables, the predictive accuracy will be higher than if only a few variables where used.

Limitations that may be faced are that many predictive models will require numerical values for all columns used. Thus, there will be a need for data conversions of some method to convert the values.

Challenges that one may face is the possibility that to better train the model some degree of oversampling may be needed to ensure that the model has a proper distribution of data to work with.

Future used of this kind of model can be used to predict business concerns like churn so that efforts could be made to keep the customers before leaving. Also, one could predict loan default to better assess the kind of interest rates to charge various client groups.

Ethically, it will be important not to consider any columns like race or culture as they could introduce bias in what the actual outcome of the model is representing.

Potential questions that may be asked:

1) What's the benefit to being able to predict who may become a customer?

2) Would this exploration be of value somewhere else?

3) Is there bias such as race or culture taken into consideration to exclude?

4) How accurate will the data model need to be to be useful?

5) Is this model and method only practical for banking?

6) What kind of challenges do you see having with the data?

7) What kind of measures would be taken to prevent revealing sensitive data?

8) Will any adjustments need to be made to the data?

9) How will the data be standardized or normalized?

10) What kind of negative results could result from this project?

Appendix

At this time there are no references and all illustrations were created by me in Power BI.