Bank Marketing (Campaign) --- Data Science

Data Glacier

Your Deep Learning Partner

Group Name

The Data Team.

Final Report Document

Team member details:

1. Fawzi El Khatib, fawzi_khatib@hotmail.com, Lebanon, Data Glacier, Data Science.
2. Nigel Chitere, tawanda.nigel@gmail.com, Zimbabwe, Rocapply, Data Analysis.
3. Malipalema Khang, malipalema@gmail.com, South Africa, University of the Witwatersrand, Data Science.
4. Arthur Mupfumira, artmupf@gmail.com, Zimbabwe, Data Glacier, Data Science.

Github repository link:

https://github.com/Fawzikh/DG-BankMarketing-Campaign

Project description

The goal of the project was to develop a model that would aid our client ABC Bank identify which clients were likely to be interested in their term deposit product before actually introducing this product into their customer base without prior research and client targeting.

Project deliverables:

The objectives of the project were as follows

1. Identify and shortlist target specific clients.
2. Build a model that can predict which clients are likely to say ‘YES’ to subscribing to the term deposit product.
3. **Save resources and the time** of the client, meaning it should be effective and accurate in its findings and predictions.

Business understanding.

In order to actually get a better picture of how this product was to be relevant and to who it would be beneficial the following questions had to be addressed

1. What is a term deposit?
2. Who is likely to be in need of this product?
3. Is this a seasonal or permanent product feature?
4. What have been the previous reactions of clients to new features introduced by the bank and what prior mechanisms or campaigns were implemented in order to ensure the success of these features?
5. Who is the target audience for this product? Any specific age group, any specific?

This gave us a better understanding and perspective into not only the client, but their line of business.

Hence the responses we got to the following question in the initiation phase included:

1. A term deposit is a fixed-term investment that includes the deposit of money into an account at a financial institution. Term deposit investments usually carry short-term maturities ranging from one month to a few years and will have varying levels of required minimum deposits.The investor must understand when buying a term deposit that they can withdraw their funds only after the term ends.

This allowed us to actually understand what product the bank was actually proposing to introduce into the market.

2. Consequently this helps to give better insights into who would be most suitable for this product and ideally that would be individuals with

   1. long term savings,
   2. stable stream of income

   Hence we are considering either people in their pension/ retirement years or those currently within the working age ratio.There may also be other major factor to consider such as current financial commitments e.g those with mortgages and loans, bonded to tertiary grants ne it students, or parents paying for their students etc

This allowed us to have a rough idea of what to look for in the data so as to support the assumption.

3. Is this a seasonal or permanent product feature?

   This would help to establish the marketing span of the product, the time, resources and effort to be put into the project. Basically is this a beta project or a long term planned project. For instance will the product have different phases so as to actually evaluate the successive growth of the initiative based on continuous evaluation or is it a one time off initiative which the bank is trying out. It could perform below expectation on the first attempt but after revision and improvement based on first attempt review and consultation the product may out-perfom itself.

This question was vital in helping to synthesize the question on whether the product was meant not only to be an ongoing campaign but one which was also based on certain periods of the year e.g fall, summer, festive season, beginning of the year etc.

4. What have been the previous reactions of clients to new features introduced by the bank and what prior mechanisms or campaigns were implemented in order to ensure the success of these features?

   This information will most likely be typified by the current % or ratios of current customers subscribing to other products, e.g home loans, bursary loans etc. This will give an indication into the general propensity of the bank's clients towards an inclination to the bank's products and services. It will also help to establish the relationship between the bank and its clients. Is it popular, does it have a large client base and following, client retainability, ease of conversion etc.

This question was vital as it allowed us to narrow down the target audience based on the successful relationship between the bank and its clients. It was also satisfactorily answered.

5. Who is the target audience for this product? Any specific age group, any specific?

From the description above, the target market definitely seems to be individuals with less financial commitments e.g loans and mortgages and yet have disposable income. Hence the target audience may involve Pensioners, Middle Aged Single and gainfully employed clients ,etc.

This essentially was the main draw point that then allowed us to build the model based on the identified target audience .

Time we had to complete the project.

To complete this project, we effectively had **1 month**. From the time we formulated and submitted our group on the **15 of April**, we were given until **15 May** to submit the final project. As such the project was undertaken in the 30 days from initiation, planning, executing to closing.

The resources we used to complete the project.

In order to achieve the objectives set out for the project, our team had to use several resources including the following below:

For communication:

Telegram, google meet, gmail and messenger.

For documentation:

We mainly used the google applications as this allowed all data to be in sync hence everyone could have access to it. Thus we used google docs, google sheets, google slides

For coding:

Our main utility used here was Jupyter Notebook. Considering that it integrated with python, this made it a perfect application as we used python for the actual model building, analyzing and visualizing the data using packages such as seaborn, pandas, matplotlib, Scikit-learn etc.

Data understanding.

In order to actually get a better picture of how this data would be relevant and to who it would be beneficial towards the successful completion of the projects objectives, we had to examine and answer the questions highlighted below:

To answer these questions we explored our data using **excel and jupyter notebook.**

The nature of our data:

We received raw and unprocessed data, which meant it had to be cleansed for machine learning suitability purposes.To fix this we used a variety of different techniques such as encoding, labelling, value dropping etc. The data was in Csv format.

The source of our data:

The given data is a publicly available dataset available for research, which was given to us as per our project requirements. The dataset was retrieved from

1. [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, In press, http://dx.doi.org/10.1016/j.dss.2014.03.001

2. Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology.
3.
4. P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011. EUROSIS.

And is available at:

[pdf] http://hdl.handle.net/1822/14838

[bib] http://www3.dsi.uminho.pt/pcortez/bib/2011-esm-1.txt

[pdf] http://dx.doi.org/10.1016/j.dss.2014.03.001

[bib] http://www3.dsi.uminho.pt/pcortez/bib/2014-dss.txt

Data Background.

The data we used for this project is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to assess if the product (bank term deposit) would be (or not) subscribed.

How many data sets:

The file includes **two datasets**:

1. bank-additional-full.csv with all examples, ordered by date (from May 2008 to November 2010).
    - Number of Instances: 45211 for bank-full.csv (4521 for bank.csv)
    - Number of Attributes: 16 + output attribute.
2. bank-additional.csv with 10% of the examples (4119), randomly selected from bank-additional-full.csv.
    - Number of Instances: 41188 for bank-additional-full.csv
    - Number of Attributes: 20 + output attribute.

Understanding the data features.

Combined the dataset has approximately close to 40 attributes.

These include **numerical** and **categorical** data.

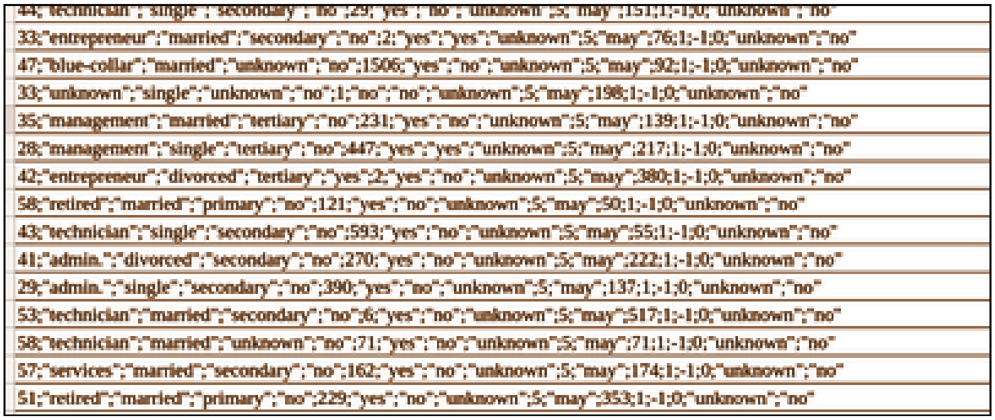
Some of the key features vital to the building of a machine learning model to predict in the decision making include the following:

1. age (numeric)
2. job : type of job (categorical: "admin.","blue-collar","entrepreneur","housemaid","management","retired","self-employed","services","student","technician","unemployed","unknown")

3. marital : marital status (categorical: "divorced","married","single","unknown"; note: "divorced" means divorced or widowed)
4. education (categorical: "basic.4y","basic.6y","basic.9y","high.school","illiterate","professional.course","university.degree","unknown")
5. default: has credit in default? (categorical: "no","yes","unknown")
6. loan: has personal loan? (categorical: "no","yes","unknown")

Problems observed in the data:

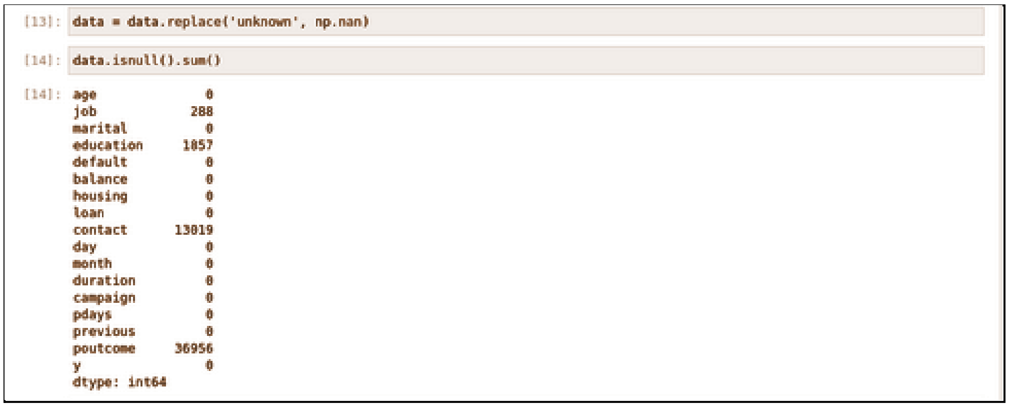
We discovered several missing values in the given data set features attributes,which were highlighted as unknown.



```
44;"technician";"single";"secondary";"no";29;"yes";"no";"unknown";5;"may";151;1;-1;0;"unknown";"no"
33;"entrepreneur";"married";"secondary";"no";2;"yes";"yes";"unknown";5;"may";76;1;-1;0;"unknown";"no"
47;"blue-collar";"married";"unknown";"no";1506;"yes";"no";"unknown";5;"may";92;1;-1;0;"unknown";"no"
33;"unknown";"single";"unknown";"no";1;"no";"no";"unknown";5;"may";198;1;-1;0;"unknown";"no"
35;"management";"married";"tertiary";"no";231;"yes";"no";"unknown";5;"may";139;1;-1;0;"unknown";"no"
28;"management";"single";"tertiary";"no";447;"yes";"yes";"unknown";5;"may";217;1;-1;0;"unknown";"no"
42;"entrepreneur";"divorced";"tertiary";"yes";2;"yes";"no";"unknown";5;"may";380;1;-1;0;"unknown";"no"
58;"retired";"married";"primary";"no";121;"yes";"no";"unknown";5;"may";50;1;-1;0;"unknown";"no"
43;"technician";"single";"secondary";"no";593;"yes";"no";"unknown";5;"may";55;1;-1;0;"unknown";"no"
41;"admin.";"divorced";"secondary";"no";270;"yes";"no";"unknown";5;"may";222;1;-1;0;"unknown";"no"
29;"admin.";"single";"secondary";"no";390;"yes";"no";"unknown";5;"may";137;1;-1;0;"unknown";"no"
53;"technician";"married";"secondary";"no";6;"yes";"no";"unknown";5;"may";517;1;-1;0;"unknown";"no"
58;"technician";"married";"unknown";"no";71;"yes";"no";"unknown";5;"may";71;1;-1;0;"unknown";"no"
57;"services";"married";"secondary";"no";162;"yes";"no";"unknown";5;"may";174;1;-1;0;"unknown";"no"
51;"retired";"married";"primary";"no";229;"yes";"no";"unknown";5;"may";353;1;-1;0;"unknown";"no"
```

Fig.1 Image showing sample data missing ‘unknown’ values anomalies in the data set.

Finding the missing/null values highlight pandas calls/functions used.



```
[13]: data = data.replace('unknown', np.nan)

[14]: data.isnull().sum()

[14]: age              0
      job            288
      marital          0
      education     1857
      default          0
      balance          0
      housing          0
      loan             0
      contact      13019
      day              0
      month            0
      duration         0
      campaign         0
      pdays            0
      previous         0
      poutcome     36956
      y                0
      dtype: int64
```

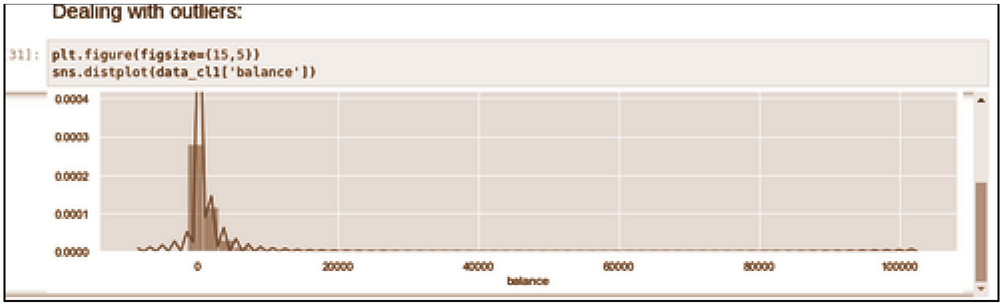
Fig.2 Image showing the ‘poutcome’ feature having the most unknown values, hence we dropped this feature given it didn't have much of a bearing on the prediction model.

What approaches are you trying to apply on your data set to overcome problems like NA value, outlier etc and why?

To address the problem above we used the following techniques class labelling(dummies), deletion and imputation techniques.

Some of the data features were discrete data such as ‘pdays, previous’, hence despite dealing with the outliers it was difficult to still have a bull curve in the output graphs.

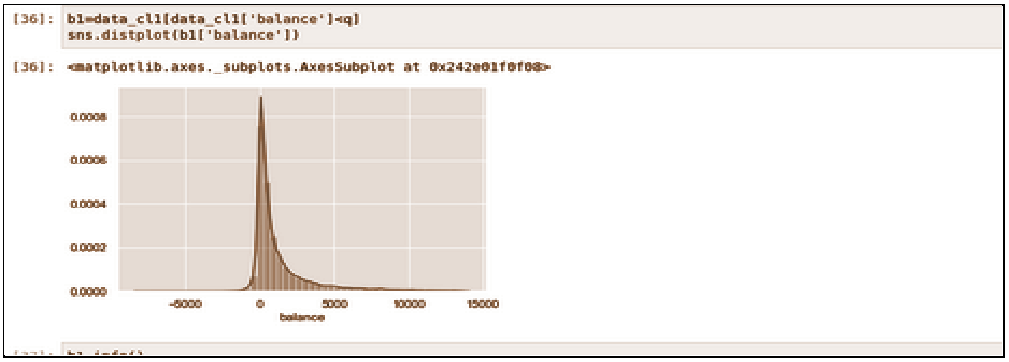
We also managed to smooth out the curves for features such as, ‘duration’, ‘balance’ ,etc.



Dealing with outliers:

```
[31]: plt.figure(figsize=(15,5))
      sns.distplot(data_cl1['balance'])
```

Fig.3 showing graphical representation of balance feature values before being worked on.



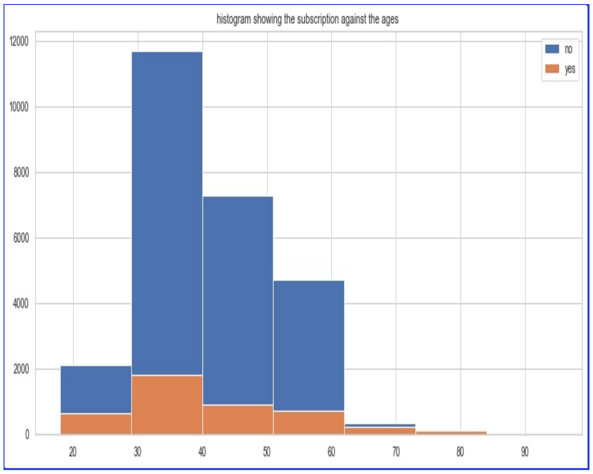
```
[36]: b1=data_cl1[data_cl1['balance']<q]
      sns.distplot(b1['balance'])

[36]: <matplotlib.axes._subplots.AxesSubplot at 0x242e01f0f08>
```

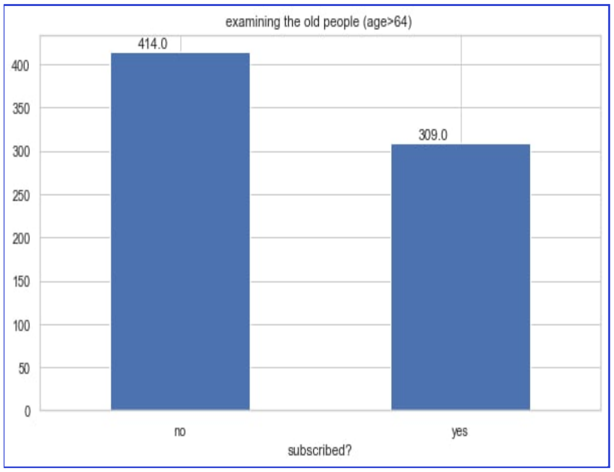
Fig.3 showing graphical representation of balance feature values before after being worked on.

Exploratory Data analysis.

In understanding each given dataset it was imperative to also establish the relations between the data features and how these were actually of any use in helping to answer or solve the problem at hand. Hence this resulted in a number of visual graphics being deduced (as will be shown in the following slides) in order to highlight the significance of some of the features within the given data sets.

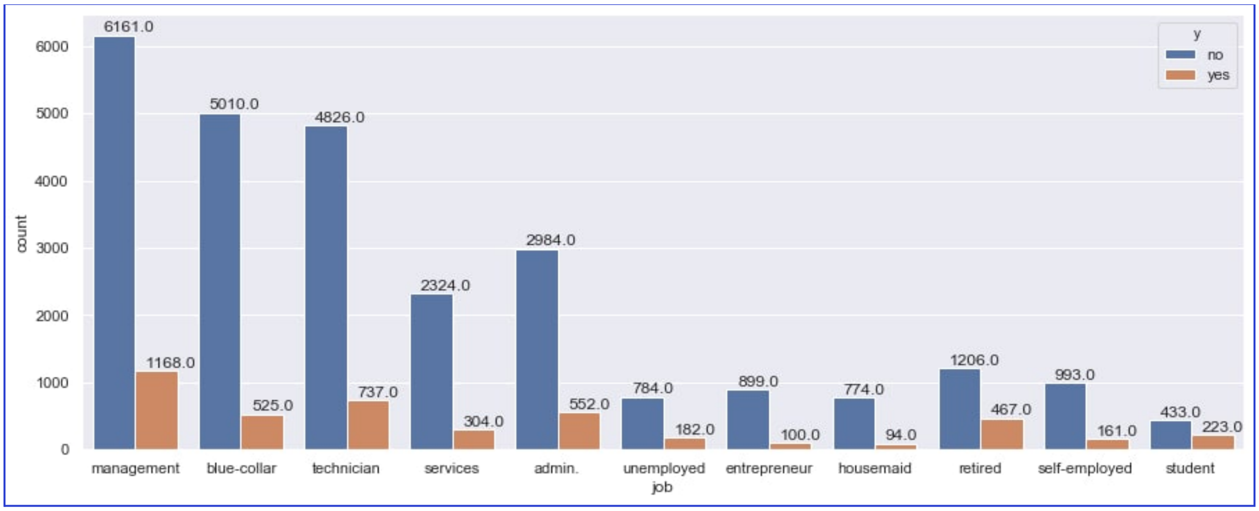
1. Data insight analysis (Age Feature Analysis).





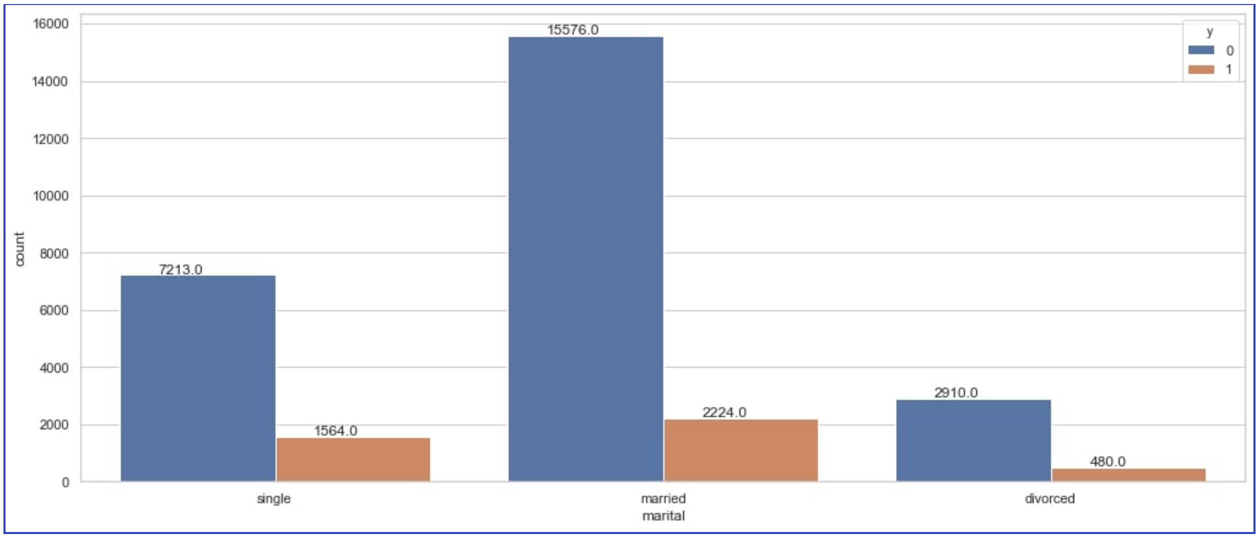
The age feature actually presented interesting insights as to which age range could afford the product. Given that the term deposit is for people with extra savings and willing to put their money away for a prolonged period of time the first group to be considered was the 'retired and working class' population. 43% over 64 agreed to the new product, while only 13% of the working age population said yes. Hence this allows us to narrow in our target group to the retired age group.

2. Data insight analysis (Job Feature Analysis).



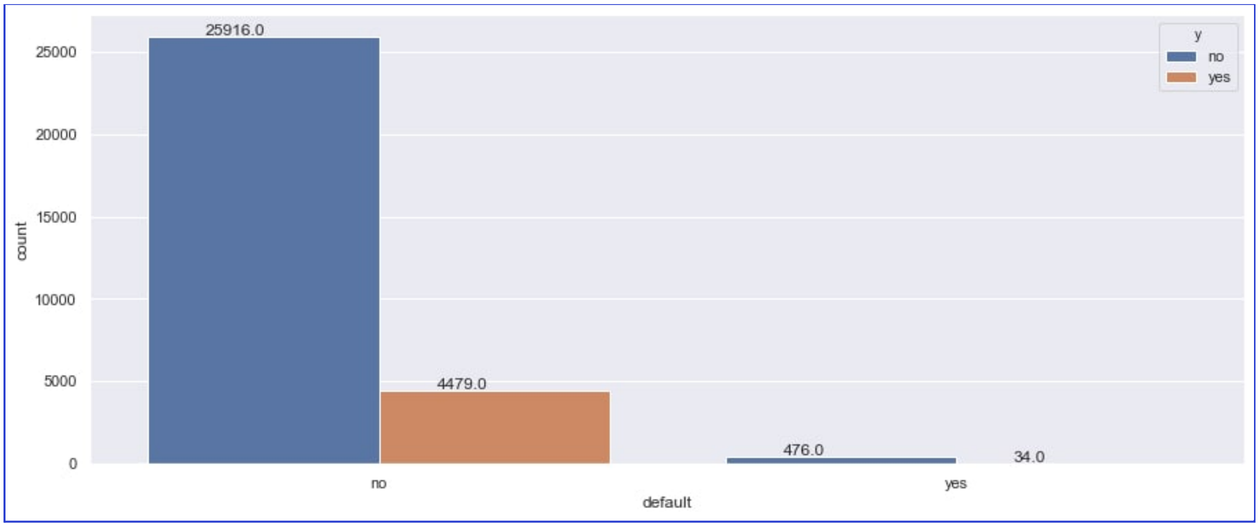
As is visible from this graph over 80% of the working population said No, however what is interesting is the ratio of yes to no for students as 34% of the students asked agreed to try out for this new campaign meaning they are a great target. Followed closely by the retired at 28% and lastly admin at 16%. This supports the previous slide that the age group over 64 years have a higher propensity to be interested in this product.

3. Data insight analysis (Marital Feature Analysis).



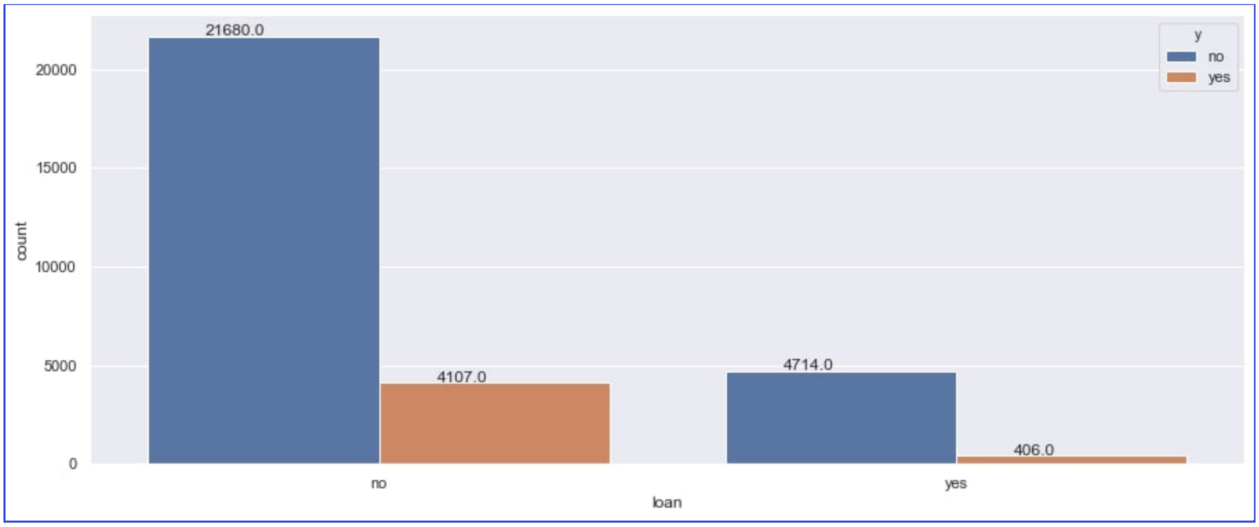
The analysis showed that the highest ratio of those who agreed were single at 22% followed by divorced at 16%. This perfectly aligns with the assumption that these social groups are likely to be interested in the term deposit product as they may not have several responsibilities such as children, school fees, joint loans or mortgages, etc. Hence these groups would be a perfect target for marketing campaigns.

4. Data insight analysis (Default Feature example).



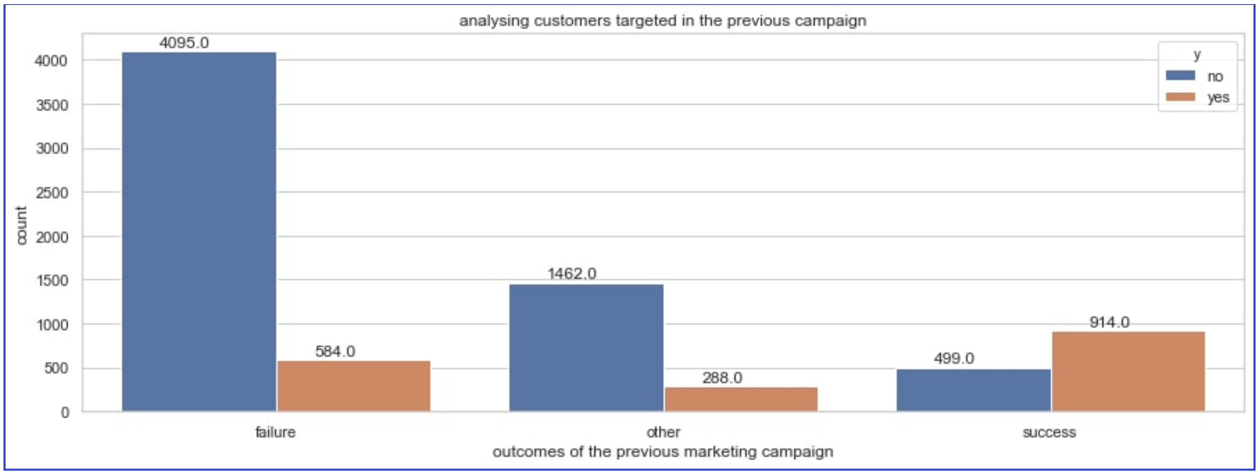
The graphical image shows the two categories of clients interest on the term deposit from those who defaulted and did not default. 19% from those who have not defaulted on previous bank commitments agreements agreed whereas only 8% of those who have defaulted said yes. This allowed us to focus on those who have **not defaulted** as their previous commitment history counts in their favour.

5. Data insight analysis (Loan Feature example).



A point of interest from this visualization is that the majority of those interested in the term deposit were clients who currently don't have any loan commitments. From a business perspective this is a positive insight as such individuals are likely not to be affected by a bad previous credit score and are less likely to default and have more money for savings as they dont have any current financial commitments with the bank.

6. Data insight analysis (Previous Campaign Feature example).



From the illustration, the largest ratio 65% of those interested in the new term fixed deposit, are clients who were successfully targeted in the previous campaigns. This could have been due to a variety of factors e.g. good customer retention, satisfactory terms and conditions etc, but more importantly it allows the marketing team to work with clients who are already keen to try new products and have a history with the bank of previously successful campaigns.

Data Insight Recommendations

After having analyzed the data above, the following recommendations were deduced:

1. The marketing campaign has to target retired people as 43% of them advised they would be interested in the new product.
2. From a marital status point, over 30% of those interested were single and divorced hence these two bins would make attractive target market points.
3. The students, retired and admin groups in the employed attribute were more keen to try out the feature therefore narrowing down the clients to focus on based on employment status.
4. Lastly with regards to previous client history consideration, it came as no brainer that clients who had not previously defaulted, did not have any current loan commitments and also those who were successfully subscribed for the previous campaigns made the majority of those interested in the new term deposit feature. Consequently backed by a good hostical standing such clients should be first for marketing consideration in this project.

Techniques we used in modelling.

In this project we are tackling a classification problem so we built four models from two different types.

The model we built was based on the .Bank-Full data set. We implemented three different logistic regression models. In the data set, we dropped attribute ‘pdays’ as it didn't not bear much impact on the model prediction, and the ‘poutcome’ feature as it had a lot of missing values.

The last model was built without standardization and duration features so as to test the general accuracy. Also we developed a **random forest model**, but without dropping the first dummy features as this model is not affected by multicollinearity.

For Regression models we know that we should prevent multicollinearity. To do so we dropped the first category of each dummy variable when creating the dummies. We specify the “ drop_first argument to True ”. Take into consideration that dummies are only compared to their respective benchmark, which is the dropped category .

Moreover, for our logistic regression model we created 3 models, 2 with standardized numerical features and one without scaling the features. The purpose of the standardization is to fix the order of magnitude issue. Thus, prevent our model from learning patterns because of the magnitude of numbers.

Removing outliers :(This is a common step for both types of models we are going to use.) for Numerical features

1. get the quantile (‘balance’, ‘duration’, ‘campaign’). Using “ “.quantile(method, so we deal with the skewness to the right(problem with the high end values). Or “ “.quantile(if the problem with low end values.
2. Filtering the data on observations that are lower or higher than the respective quantile.

II. Transformations of the object features : (In this step we used two techniques)

1. for Binary (‘yes’ or ‘no’) features : we mapped the data using “ “.map ” method. Where we mapped yes 1 and no 0. This is what we did for “housing”, “default”, and “loan” features
2. for Categorical features the values of each feature here represent categories that are equally meaningful. For the purpose of making quantitative analysis, to add numeric meaning to our categorical nominal values, we used dummy variables pd.get_dummies method. This is how we transformed 'job’, 'marital’, and 'contact’ features.

Model findings

Our developed model had an accuracy of a metric accuracy score of 86% for the random forest model. The trained model was giving results in approximately 3 secs at most after having fed the model a total of 22475 instances of data. As such this meant the model calculation was quite fast.

Problems encountered

As with any project it is critical to also acknowledge the challenges faced. For our team, some of them included:

1. Same accuracy. For the model without the duration feature and standardization.

   After having tried to improve the accuracy of the model by dropping the data features to test the effect of each dropped feature per given time, we still obtained the same model prediction accuracy which was 85% hence this was a challenge for us as we failed to identify what was hindering the model from predicting a higher prediction result.

2. Teamwork and cohesion.

   As with any project, the ability to work together and jel as a team is one you have to overcome in order to successfully complete a given project. In our team, the tasks were made difficult by the non availability of other members which would have lessened the work for each of us and helped us to come up with an even better project result.

3. Time constraints

   Given that some in the team were working professionals, it was difficult to work around the clock so as to come up with workable times for everyone in the team. Although it was possible, it required a lot of sacrifices late night working and constant communication within the team. Furthermore, the time allocated for the project 30 days given the load was not enough, however our team made it work and was satisfactorily able to complete the project objectives as expected.

What we learnt from this project:

This project allowed us to learn a lot of different skills both soft and technical skills. As with any project, we had to understand the problem on its merits. We had to learn the terms, language and references used in the banking industry so as to understand how to deal with the problem. This project also helped us to develop our analytical and technical skills, together with team collaboration and problem solving.

Overall Conclusion.

The project was an overall success as we were able, as highlighted in the topical issues discussed above, to achieve the project deliverables with resounding success. We deployed a model that had above 85% accuracy in terms of prediction and we were able to identify the target audience on which this model should focus on so as to help the client attain the desired results.

Acknowledgements

Special appreciation to this project contributors:

1. Fawzi El Khatib, fawzi_khatib@hotmail.com, Lebanon, Data Glacier, Data Science.
2. Nigel Chitere, tawanda.nigel@gmail.com, Zimbabwe, Rocapply, Data Analysis.

Special appreciation also to the Data Glacier team for their mentorship and guidance which allowed us to successfully complete this project and helped us improve and gain much needed experience in the world of Data Analytics and Science.