Project details - classification

Background:
You are working as a risk analyst with a bank. Apart from the other banking and loan services, the bank also provides credit card services which is a very important source of revenue for the bank. The bank wants to understand the demographics and other characteristics of its customers that accept a credit card offer and that do not accept a credit card.
Usually the observational data for these kinds of problems is somewhat limited in that often the company sees only those who respond to an offer. To get around this, the bank designs a focused marketing study, with 18,000 current bank customers. This focused approach allows the bank to know who does and does not respond to the offer, and to use existing demographic data that is already available on each customer.

Objective: The task is to build a model that will provide insight into why some bank customers accept credit card offers. There are also other potential areas of opportunities that the bank wants to understand from the data.
Your senior management has also posted these other questions that will help them better understand their customers.

Data:
The data set consists of information on 18,000 current bank customers in the study. These are the definitions of data points provided:

- **Customer Number**: A sequential number assigned to the customers (this column is hidden and excluded – this unique identifier will not be used directly).
- **Offer Accepted**: Did the customer accept (Yes) or reject (No) the offer. Reward: The type of reward program offered for the card.
- **Mailer Type**: Letter or postcard.
- **Income Level**: Low, Medium or High.
- **#Bank Accounts Open:** How many non-credit-card accounts are held by the customer.
- **Overdraft Protection**: Does the customer have overdraft protection on their checking account(s) (Yes or No).
- **Credit Rating**: Low, Medium or High.
- **#Credit Cards Held**: The number of credit cards held at the bank.
- **#Homes Owned:** The number of homes owned by the customer.
- **Household Size**: Number of individuals in the family.
- **Own Your Home**: Does the customer own their home? (Yes or No).
- **Average Balance**: Average account balance (across all accounts over time). **Q1, Q2, Q3 and Q4**
- **Balance**: Average balance for each quarter in the last year

Exploring the data

We encourage you to thoroughly understand your data and take the necessary steps to prepare your data for modeling before building exploratory or predictive models. Since this is a classification model, you can use logistic regression for classification for building a model. You are also encouraged to use other models in your project including KNN classifiers, decision trees.
To explore the data, you can use the techniques that have been discussed in class. Some of them include using the describe method, checking null values, using _matplotlib_ and _seaborn_ for developing visualizations.

The data has a number of categorical and numerical variables. Explore the nature of data for these variables before you start with the data cleaning process and then data pre-processing (scaling numerical variables and encoding categorical variables).
For the target variable (Offer accepted – Yes/No), it is also important to check the data imbalance ie the number of people who responded with a yes vs the number of people who responded with a no.
You will also use tableau to visually explore the data further. You will deep dive in the data for customers who accepted the offer vs the customers who did not and check their characteristics. For e.g., we select the **Yes** level in **Offer Accepted** and then examine the distribution of accepted offers across the other variables in our data set and similarly for people who did not accept the offer.

Model

Use different models to compare the accuracies and find the model that best fits your data. You can use the measures of accuracies that have been discussed in class. Please note that while comparing different models, make sure you use the same measure of accuracy as a benchmark.

Mid-bootcamp project deliverables

You should maintain a separate GitHub repo for this project with the following files:

- `Readme.md` - This markdown will explain the data analysis workflow including the problem statement/ business the objective, data extraction, data wrangling, etc. Here you should explain the business analytic approach you used to solve the problem. Please be detailed in explaining the steps you followed. It is important to keep in mind that the document is written for the readers, who may or may not have the technical expertise with Python/SQL/Tableau.
- Python File - It can be either uploaded as a `.ipynb` file (Jupyter notebook) or `.py` file. The Python code should be well documented with comments, explaining the code, EDA operations, logic used - especially with data cleaning operations, and any assumptions followed in the model.
- Dataset/datasets (provided to you)
- Tableau workbook
- File containing SQL queries

** You are provided with the rubrics that will be used to evaluate the projects. Please go through the document for more details on the specificities for different files.

Some other tips

- Pay attention to the naming convention: organize the files in folders with appropriate names
- Do not include code snippets in the `Readme.md` file
- Explain the business insights and the regression/classification model results
- Explain the future score of work
- Make daily commits to the repo

SQL Questions - Classification

(Use sub queries or views wherever necessary)
1. Create a database called `credit_card_classification`.
2. Create a table `credit_card_data` with the same columns as given in the csv file. You can find the names of the headers for the table in the `creditcardmarketing.xlsx` file. Use the same column names as the names in the excel file. Please make sure you use the correct data types for each of the columns.
3. Import the data from the csv file into the table. Before you import the data into the empty table, make sure that you have deleted the headers from the csv file. (in this case we have already deleted the header names from the csv files). To not modify the original data, if you want you can create a copy of the csv file as well. Note you might have to use the following queries to give permission to SQL to import data from csv files in bulk:

```sql
SHOW VARIABLES LIKE 'local_infile'; -- This query would show you the status of the variable ‘local_infile’. If it is off, use the next command, otherwise you should be good to go

SET GLOBAL local_infile = 1;
```

4. Select all the data from table `credit_card_data` to check if the data was imported correctly.
5. Use the _alter table_ command to drop the column `q4_balance` from the database, as we would not use it in the analysis with SQL. Select all the data from the table to verify if the command worked. Limit your returned results to 10.
6. Use sql query to find how many rows of data you have.
7. Now we will try to find the unique values in some of the categorical columns:

   - What are the unique values in the column `Offer_accepted`?
   - What are the unique values in the column `Reward`?
   - What are the unique values in the column `mailer_type`?
   - What are the unique values in the column `credit_cards_held`?
   - What are the unique values in the column `household_size`?

8. Arrange the data in a decreasing order by the `average_balance` of the house. Return only the `customer_number` of the top 10 customers with the highest `average_balances` in your data.
9. What is the average balance of all the customers in your data?
10. In this exercise we will use `group by` to check the properties of some of the categorical variables in our data. Note wherever `average_balance` is asked in the questions below, please take the average of the column `average_balance`:

   - What is the average balance of the customers grouped by `Income Level`? The returned result should have only two columns, income level and `Average balance` of the customers. Use an alias to change the name of the second column.
   - What is the average balance of the customers grouped by `number_of_bank_accounts_open`? The returned result should have only two columns, `number_of_bank_accounts_open` and `Average balance` of the customers. Use an alias to change the name of the second column.
   - What is the average number of credit cards held by customers for each of the credit card ratings? The returned result should have only two columns, rating and average number of credit cards held. Use an alias to change the name of the second column.
   - Is there any correlation between the columns `credit_cards_held` and `number_of_bank_accounts_open`? You can analyse this by grouping the data by one of the variables and then aggregating the results of the other column. Visually check if there is a positive correlation or negative correlation or no correlation between the variables.
   You might also have to check the number of customers in each category (ie number of credit cards held) to assess if that category is well represented in the dataset to include it in your analysis. For eg. If the category is under-represented as compared to other categories, ignore that category in this analysis

11. Your managers are only interested in the customers with the following properties:

   - Credit rating medium or high

- Credit cards held 2 or less
- Owns their own home
- Household size 3 or more

For the rest of the things, they are not too concerned. Write a simple query to find what are the options available for them? Can you filter the customers who accepted the offers here?

12. Your managers want to find out the list of customers whose average balance is less than the average balance of all the customers in the database. Write a query to show them the list of such customers. You might need to use a subquery for this problem.
13. Since this is something that the senior management is regularly interested in, create a view called `Customers__Balance_View1` of the same query.
14. What is the number of people who accepted the offer vs number of people who did not?
15. Your managers are more interested in customers with a credit rating of high or medium. What is the difference in average balances of the customers with high credit card rating and low credit card rating?
16. In the database, which all types of communication (`mailer_type`) were used and with how many customers?
17. Provide the details of the customer that is the 11th least `Q1_balance` in your database.

Tableau - Classification

In this part of the project you will work with the data set `creditcardmarketing.xlsx' use Tableau to answer the questions below. Make a separate sheet for every question:

Tableau Questions:

1. Convert the necessary measures to dimensions (the variables that are categorical in nature). When you use a separate sheet for this question, add a note in that sheet on which columns were changed
2. Check the imbalance in the dataset by looking at the number of people who accepted the offer vs. people who did not accept the offer. Add the counts as labels on the plots
3. Do a quick table calculation on the previous plot to check percentage of total for both _yes_ and _no_.
4. Now we will try to analyze certain characteristics / the differences between the people who accepted the offer vs people who did not accept the offer. Use the same sheet for plots below.
   - Plot average Q1 balance vs Offer Accepted. Provide the values of averages as labels.
   - Plot average Q2 balance vs Offer Accepted. Provide the values of averages as labels.
   - Plot average Q3 balance vs Offer Accepted. Provide the values of averages as labels.
   - Plot average Q4 balance vs Offer Accepted. Provide the values of averages as labels.
5. We saw all the four plots together on the same sheet. The plots should have the same format for numbers (number of decimal places here). Do you observe any trend here. Add a caption to provide the explanation
   - Now for all the plots, change the style of the plot from bar chart to a line chart. This could be used for a visual trend
6. Consider a similar analysis for Household Size vs average balances for each quarter. You would observe a huge jump in average balance from Q1 to Q2 for households with size 8.
   - Try and explain that jump. Hint: Check the number of records we have for such customers. Do you see any anomaly.
7. Now we want to see how some of the other features In the data might have affected responses from the people. For these we will first start by creating a cross tab. A cross tab is simply a table between two categorical features with some metric of importance filling up the table.

   - Create a cross tab between Offer Accepted and Overdraft Protection and fill the table with number of records. Do you observe any trend here?
   - Create a cross tab between Offer Accepted and Mailer Type and fill the table with number of records. Do you observe any trend here?
   - Create a cross tab between Offer Accepted and Credit Rating and fill the table with number of records. Rearrange the column credit rating from low to high. Do you observe any trend here?

8. Based on the average balance for each customer, create four buckets : Category A, Category B, Category C, and Category D. Category A is from min value to 700, 701 to 1400, 1400 to 1900, and 1900 to 3366. Check the number of observations for each of the categories.

9. Create a visually appealing dashboard to represent the information

   Some points to keep in mind while working on the tableau questions:

   a) The plots should be well labelled briefly describing the purpose of the plot
   b) Select the chart type that produces an effective outcome for a given scenario
   c) Focus audience attention on the most important data
   d) Use space, color and fonts appropriately
   e) Use correct title for the plots.
   f) Utilize formatted tooltips and descriptive titles
   g) Format the axes wherever necessary
   h) Use caption to add details wherever necessary
   i) Use appropriate level of details with labels and color coding etc.
   j) For the dashboard make sure that the information represented is clear and easy to understand. The user of the dashboard should be able to understand the purpose of the dashboard and should be able to make decisions looking at the plots presented.
   k) You can also use filters wherever appropriate to give the user the flexibility to view different information easily