UNICORN FINANCIAL

A blessing for your money!

DATA8001 Assignment – 40%

Following the recent acquisition of Grand Slam Financial by Unicorn Financial, a review of IT systems was undertaken. Management decided to integrate all systems & data sources into a single solution. As part of the restructuring, a review of the customer base is required to identify what customers are most likely to leave the bank and what customers should be targeted for other products. You have been hired to analyse data from three sources:

1. Legacy Unicorn Excel files
2. Unicorn Mongo database extract
3. Grand Slam Financials' MYSQL database extract

During the restructuring there was data corruption and owing to legacy data collection issues, the data requires considerable cleaning prior to modelling. Given the type of data & predictions required, a Logistic Regression model is proposed to predict customer churn across the 10,000 customers.

You are required to:

- **Part A) Data Cleansing in Jupyter Notebook:**
  - Load the 3 datasets into data frames (see appendix for more details)
  - Join the datasets, use the same column naming convention as the MYSQL database (see appendix)
  - Clean the datasets – see appendix for more details
  - Convert all currencies to Euro €, use the conversion rates in appendix
  - All country codes must conform to [IE, UK, USA] – see appendix
  - Save the clean dataset as a CSV file for Qlik:
    - **DATA8001_Assignment_2018_Data_Cleaned_Qlik_<STUDENT_NUMBER>.csv**
- **Part B) Summary Statistics in Jupyter Notebook:**
  - All queries below should have a corresponding matplotlib graph in the Python notebook.
    - What country has the highest exit rate of customers? What is this rate compared to other countries?
    - Are male or female customers more likely to exit the bank? Explain your answer with relevant statistics.
    - Is the number of products a customer has significant in terms of them exiting the bank?
    - What is the average exit rate for the following 5 age groups:
      - < 20yrs, 20 – 29 ,30 – 39, 40 – 59, > 60 yrs
    - Are longer term customers (> 4 years) more or less likely to exit the bank?

- **Part C) Data Visualisation in Qlik Sense:**
  - Visualise the clean dataset to enable Unicorn Financial to view the queries from Part B.
  - The Qlik application should be user friendly showing the data in a clean, uncluttered & intuitive manner.
  - Max 8 graphs across no more than 3 sheets. All sheets should be neatly presented with minimum number of graphs required to convey the detail required.
  - It's up to you to choose the best visual representation of your work to tell your story.
  - **NOTE:** You can also include these graphs or if you prefer to use *matplotlib* graphs in python in your final report & presentation.
- **Part D) Data Modelling in Jupyter Notebook**

  NOTE: Always train your model on the training set and predict using the test set! – use a 70% training set.
  - Using the clean dataset, create a Logistic Regression model of the customer data to complete the following:
    - Develop a model to predict what customers are most likely to leave the bank.
    - What is the best individual feature to predict a customer exiting the bank?
    - What are the best features (if they exist) to predict customer churn?
    - How does your model predictions compare to the statistics from Part B? Discuss.
  - **NOTE:**
    - Use a test set size of 30% & a random seed of 2018 when splitting the data

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=2018)
```

UNICORN FINANCIAL

A blessing for your money!

Files to be uploaded:

- **IPython Notebook**: containing all you python work (data ingestion, joining, cleansing, summary statistics, logistic regression modelling)
  - **DATA8001_Assignment_2018_Python_<STUDENT_NUMBER>.ipynb**
- **Qlik**: Your .QVF file (usually saved in: *Documents\Qlik\Sense\Apps*)
  - **DATA8001_Assignment_2018_Qlik_<STUDENT_NUMBER>.qvf**
- **Project Report**: MS Word / PDF report providing sufficient & succinct detail of your findings & recommendations to Unicorn Financial. Neat presentation is important.
  - **DATA8001_Assignment_2018_Report_<STUDENT_NUMBER>.doc / pdf**

Marking

This project is worth 40% of your overall mark for the DATA8001 module.

Data Cleansing	25
Data Visualisation	20
Data Modelling	25
Report	15
Presentation	15
Total Marks	**100**

Due Dates

- ALL assignments are due by Monday 16th April (Week 10) at 5pm.
- No late submissions will be accepted.

Presentation

- ALL presentations (15 marks) will be reviewed during the Week 10 lecture (7pm Tuesday 17th April).
- Each presentation is 10 minutes in duration (incl. 2 min for questions).
- ALL class members are expected to attend unless otherwise agreed.
- ALL presentation material (e.g., PowerPoint etc.) must be submitted by 5pm on the day of the presentation. Any special requests (i.e., software etc.) must be agreed beforehand, instructor laptop is the only laptop being used for all presentations.
- Recommended that each presentation is no more than 4/5 slides and allow 2mins for questions.

UNICORN FINANCIAL

A blessing for your money!

Appendix

Data Source Headings & Definitions

SQL DB **	Excel Files	Mongo DB	Description
customer_id	Id	Cust_ID	Customer ID
surname	Surname	Last Name	Customer Surname
credit_score	Credit Score	Score	Customer Credit Score
geography	Location	Country	Customer Country
gender	Gender	Gender	Customer Gender
age	Age	Age	Customer Age
tenure	Years in Employment	Current Tenure	Number of years the customer has been with the bank
balance	Current Balance	Balance	Customers current balance
num_products	Product Count	Products	The number of products the customer has taken out
credit_card	Has CC	Credit Card	Does the customer have a credit card?
active_status	Current Status	Status	Is the customer active?
salary_est	Current Salary	Estimated Salary	What is the customers estimated salary
exited	Still In Company	Existing Employee	Is the customer still with the bank or have they left?

** Use the SQL DB naming convention in the new clean dataset to be loaded into Qlik.

Mysql database table name = data8001_assignment_2018

Binary Naming Conventions

Binary Value	Description
0	**Male, M, No, False, Inactive**
1	**Female, F, Yes, True, Active**

Country Codes

Ireland	United Kingdom	North America
IE	**UK**	**USA**

Exchange Rates

EURO	USD	GBP
€1	**$1.22870**	**£0.891355**