Business Report

Contents

Executive Summary 2
Introduction 2
Dataset Requirements 2
  Overview 2
  Ethical Consideration 4
Database Engine / Design 4
  Engine 4
  Data Relationship Model 5
Software Development Methodology 5
  Overview 5
  Website Team 5
  Mobile App Team 6
  Machine learning Team 6
Software and hardware requirements 7
  Server-side software 7
  Server-side hardware 7
Costing & Staffing 7
  Cost table 7
  Staff Cost 8
Legal Considerations 8
  Data Protection Act 8
  Copyright and Licence 8
References 9
  Software References 9
  Hardware References 9

Executive Summary

The project contains a website and mobile application that will use a machine learning model to generate shopping lists and advertise products based on the user's information. The project is estimated to take 4 months to develop and cost £113,132.14 over 5 years including staff costs for development and software licencing. There are 3 teams in total for the project. Website, Machine Learning and Mobile Application. In-house hardware is part of the cost to set up space for user accounts, product listings and storage of the machine learning data and model. The legalities and ethics have been considered for the dataset used and processed user data.

Introduction

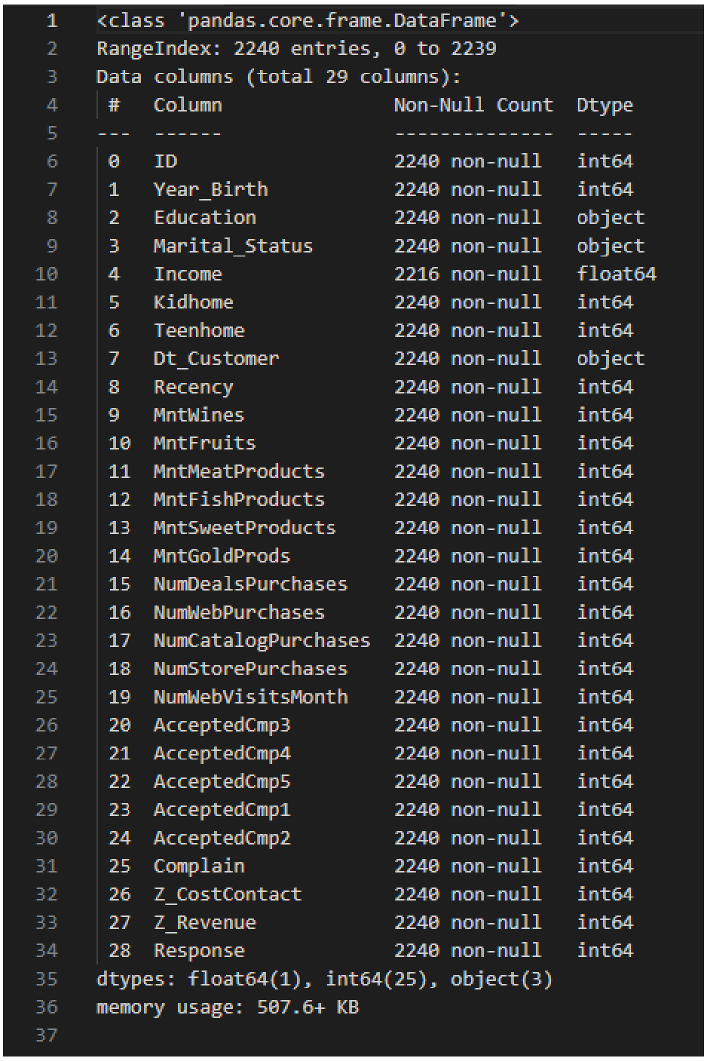
The project will consist of developing and deploying a mobile application and website that customers can access with their own accounts to get personalised shopping lists and recommended products based on a few parameters given when they make an account.

The product suggestion and shopping list generation will use machine learning to give those recommendations through training based on a relevant dataset. The website will be able to create and access user accounts, display product listings and allow them to be bought, view receipts and display active promotion campaigns or deals. The mobile application will mirror the website functionality but also allow for users to get their own barcode which can be used to store digital receipts from purchases in-store.

Dataset Requirements

Overview

The data to be used for developing the machine learning model will be from a Customer Personality Analysis dataset used from Kaggle (Patel, 2021). This contains 2240 rows and 29 columns in its raw form and is in a csv file format. A list of columns is shown below.

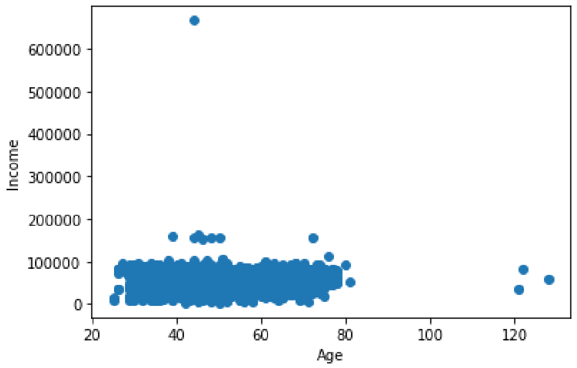


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 29 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ID                   2240 non-null   int64
 1   Year_Birth           2240 non-null   int64
 2   Education            2240 non-null   object
 3   Marital_Status       2240 non-null   object
 4   Income               2216 non-null   float64
 5   Kidhome              2240 non-null   int64
 6   Teenhome             2240 non-null   int64
 7   Dt_Customer          2240 non-null   object
 8   Recency              2240 non-null   int64
 9   MntWines             2240 non-null   int64
 10  MntFruits            2240 non-null   int64
 11  MntMeatProducts      2240 non-null   int64
 12  MntFishProducts      2240 non-null   int64
 13  MntSweetProducts     2240 non-null   int64
 14  MntGoldProds         2240 non-null   int64
 15  NumDealsPurchases    2240 non-null   int64
 16  NumWebPurchases      2240 non-null   int64
 17  NumCatalogPurchases  2240 non-null   int64
 18  NumStorePurchases    2240 non-null   int64
 19  NumWebVisitsMonth    2240 non-null   int64
 20  AcceptedCmp3         2240 non-null   int64
 21  AcceptedCmp4         2240 non-null   int64
 22  AcceptedCmp5         2240 non-null   int64
 23  AcceptedCmp1         2240 non-null   int64
 24  AcceptedCmp2         2240 non-null   int64
 25  Complain             2240 non-null   int64
 26  Z_CostContact        2240 non-null   int64
 27  Z_Revenue            2240 non-null   int64
 28  Response             2240 non-null   int64
dtypes: float64(1), int64(25), object(3)
memory usage: 507.6+ KB
```

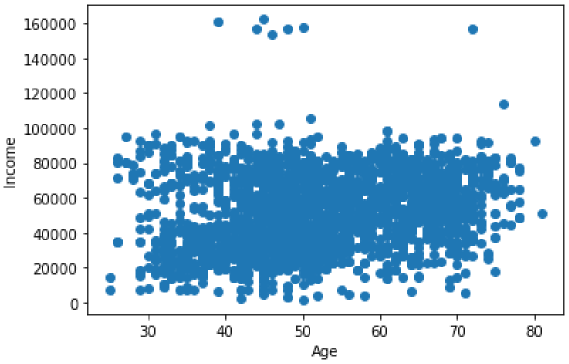
Data such as amount spent on food items, combined with income, number of children and responsiveness to advertising campaigns can be used to create a machine learning model for new users to get relevant adverts and customised shopping list recommendations on the website or mobile app.

To clean and organise the data, the dataset should be split into 4 separate data frames based on people, products, promotions, and place. Features are created on the people data frame called ‘Age’, ‘TotalSpent’ and ‘NumChildren’ to better visualise what data is there. The ‘Age’ feature was created by taking the Current system times year then subtracting the year of birth from the dataset. This was then used to create a simple scatter plot to show ‘Age’ against ‘Income’.

Before limits



After Limits



This showed that there were some outliers in the data that would affect the training of a machine learning model. The Income column will need to have any rows over £500,000 to be removed and for the Age to be limited to under 100 years old. These modifications will need to be applied before the dataset is stored in a database and will reduce the total usable rows to 2212. This will also need to be split into a train and test set before using the data.

Ethical Consideration

For the dataset to be safely used, there needed to be some criteria set. The dataset could not contain any of the following:

- Personal Identifiers – Names, Specific locations.
- Biased, Gendered information
- Racial bias

For protected Characteristics such as Age, Disability, Race, Sexual orientation if present, they cannot be used to discriminate against that individual or be modified to introduce a bias under the Equality act 2010.

The users can create and modify their accounts and have the right to be informed how their data is being used. In the case of this project, their personal data must be protected as it contains bank details and personal identifiers.

Database Engine / Design

Engine

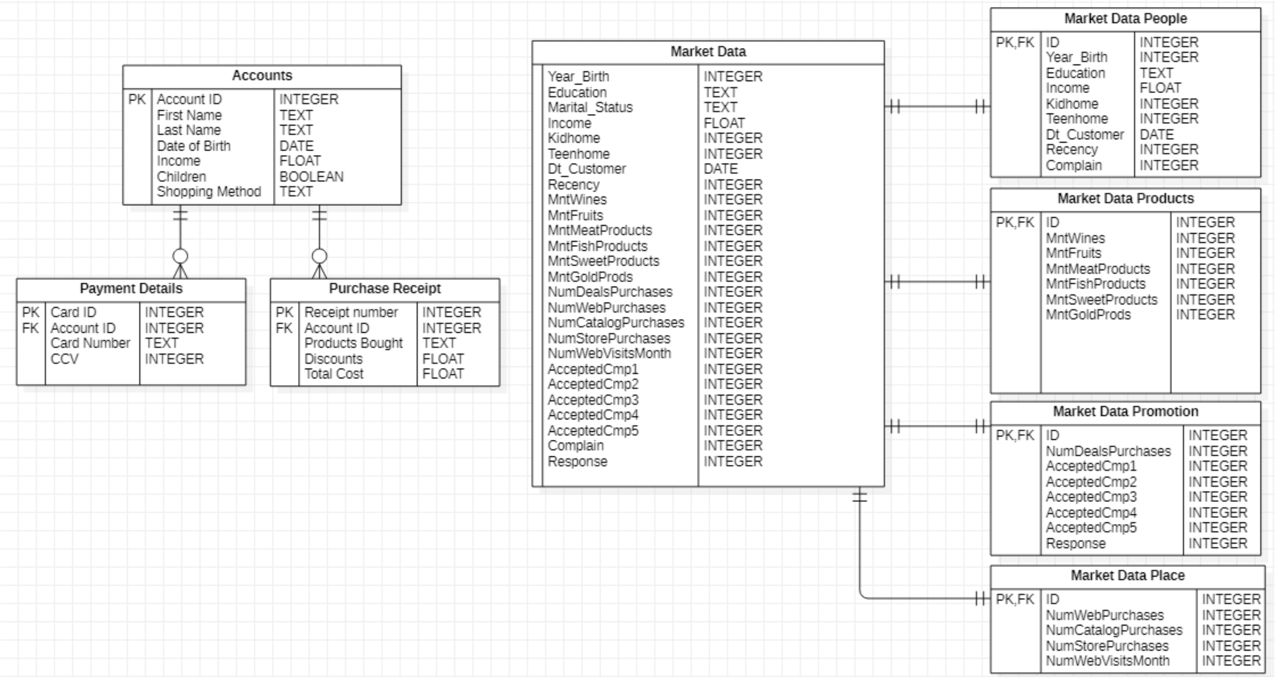
The possible database engine options are:

- Oracle - £35582.26 per unit (socket x core x core factor)
- MySQL - £3745.50 per 4 socket server.
- MS SQL Server - £10298.63 per 2 cores

From the above options, the cheapest option would be MySQL. This would allow for scalability up to 4 sockets per server before more licenses are needed.

For the database to be easily accessed from one place, an ORM (Object-relational mapping) will be used to connect to the storage server. As we would use MySQL, we have access to free-open source ORM options such as SQLAlchemy, DapperORM and Sequelize.

Data Relationship Model



Accounts		
PK	Account ID	INTEGER
	First Name	TEXT
	Last Name	TEXT
	Date of Birth	DATE
	Income	FLOAT
	Children	BOOLEAN
	Shopping Method	TEXT

Payment Details		
PK	Card ID	INTEGER
FK	Account ID	INTEGER
	Card Number	TEXT
	CCV	INTEGER

Purchase Receipt		
PK	Receipt number	INTEGER
FK	Account ID	INTEGER
	Products Bought	TEXT
	Discounts	FLOAT
	Total Cost	FLOAT

Market Data		
	Year_Birth	INTEGER
	Education	TEXT
	Marital_Status	TEXT
	Income	FLOAT
	Kidhome	INTEGER
	Teenhome	INTEGER
	Dt_Customer	DATE
	Recency	INTEGER
	MntWines	INTEGER
	MntFruits	INTEGER
	MntMeatProducts	INTEGER
	MntFishProducts	INTEGER
	MntSweetProducts	INTEGER
	MntGoldProds	INTEGER
	NumDealsPurchases	INTEGER
	NumWebPurchases	INTEGER
	NumCatalogPurchases	INTEGER
	NumStorePurchases	INTEGER
	NumWebVisitsMonth	INTEGER
	AcceptedCmp1	INTEGER
	AcceptedCmp2	INTEGER
	AcceptedCmp3	INTEGER
	AcceptedCmp4	INTEGER
	AcceptedCmp5	INTEGER
	Complain	INTEGER
	Response	INTEGER

Market Data People		
PK,FK	ID	INTEGER
	Year_Birth	INTEGER
	Education	TEXT
	Income	FLOAT
	Kidhome	INTEGER
	Teenhome	INTEGER
	Dt_Customer	DATE
	Recency	INTEGER
	Complain	INTEGER

Market Data Products		
PK,FK	ID	INTEGER
	MntWines	INTEGER
	MntFruits	INTEGER
	MntMeatProducts	INTEGER
	MntFishProducts	INTEGER
	MntSweetProducts	INTEGER
	MntGoldProds	INTEGER

Market Data Promotion		
PK,FK	ID	INTEGER
	NumDealsPurchases	INTEGER
	AcceptedCmp1	INTEGER
	AcceptedCmp2	INTEGER
	AcceptedCmp3	INTEGER
	AcceptedCmp4	INTEGER
	AcceptedCmp5	INTEGER
	Response	INTEGER

Market Data Place		
PK,FK	ID	INTEGER
	NumWebPurchases	INTEGER
	NumCatalogPurchases	INTEGER
	NumStorePurchases	INTEGER
	NumWebVisitsMonth	INTEGER

Software Development Methodology

Overview

The project will contain 3 development areas. The mobile application, website and Machine learning. Each area should have its own team following a software development life cycle to efficiently produce features in each time frame. Possible frameworks for development could be, Scrum, Feature-Driven Development, Waterfall and Agile.

Website Team

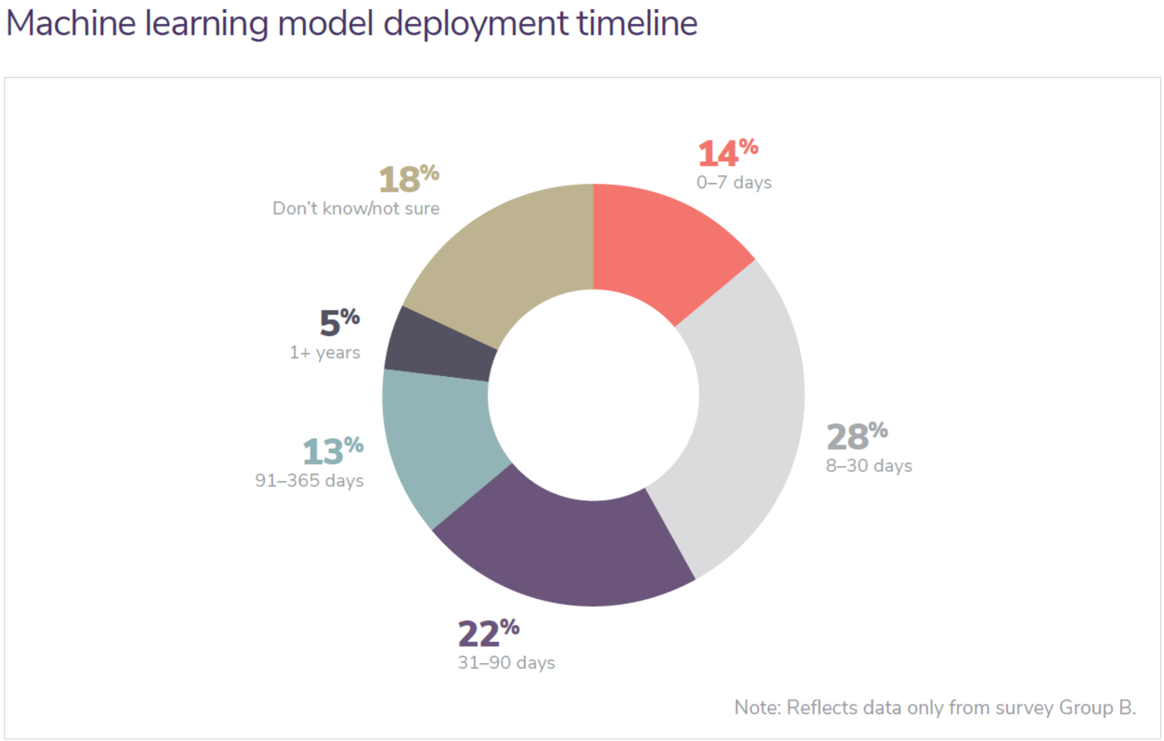
The website will need front-end and back-end development to function. The front end will consist of html, CSS, and JavaScript. The back-end will consist of python collecting user's data, building a data frame or adding to it, and using an ORM to link it to a database so it is updated in real-time. As the website will have more tasks involved to complete, taking a scrum approach will equally produce functionality in key areas. The most efficient team size for scrum is 3-5 people over short lengths (Subhash, 2020). Breaking each aspect into chunks will improve efficiency and reduce the chance of scope creep. The average time to develop a web app is at least 3 months (Didenko, 2021) which will be the aim.

Mobile App Team

Like the website, the app can be developed using html, CSS and JavaScript. The app will be compatible with both Android and IOS. The initial development will start with Android and then modified to work with the IOS API. Making them native and not hybrid will greatly improve performance on each operating system (IBM, 2020). The mobile app will function mainly from existing features produced when making the website, so an ideal approach would be Feature-Driven Development to convert between the website code to Android and then IOS API. Feature-Driven is very similar to scrum except it has longer development lengths and is more attuned to a lightweight waterfall method.

Machine learning Team

The machine learning team will follow a data science specific life cycle when creating the model. The team will also be responsible for the cleaning and usage of the dataset. The team will use a notebook-based approach. The computing power for training the deep learning model will be outsourced as it will be cheaper than purchasing hardware. The deep learning model will be trained with Random Forrest generation and decision trees to produce an accurate result over multiple iterations.



Pie Chart Source: Algorithmia's "2020 State of Enterprise ML". (Algorithmia, 2020)

The average time to develop a model is between 31-365 days so taking an average of 3 months to match the web development time will be used for calculating wages and project time.

Software and hardware requirements

Server-side software

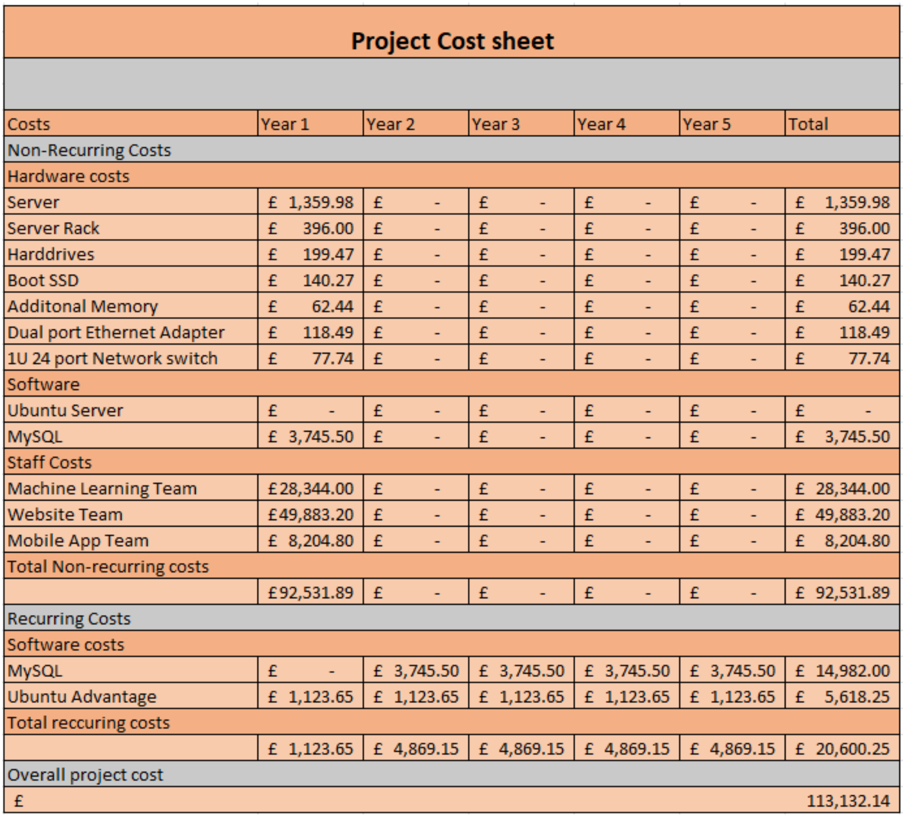
The server will run on ubuntu server which is free and open source. The database used will be MySQL combined with the SQLAlchemy ORM to interface with the database. The machine learning model and feature engineering done in python notebooks sets up the relevant pipelines for the project to be completed and function.

Server-side hardware

The server to be used is an enterprise 2U rackmount type HyperServe RMXE-2U8 customised from Novatech which has 16GB Ram, 6TB HDD in slot, and 500GB SSD for software and boot. The rackmount solution will allow for quick scalability by filling expansion slots. The base processor is a Xeon-2224 Quad-core, 4 thread processor that will allow for a low setup cost with database software. The server will be installed with Ubuntu server and be setup with a yearly subscription to Ubuntu advantage for maintenance and security. A 1U TP-link 24 port network switch will also be added to interface between the server and local computers for development. The rack solution will be an Orion 12U standing rack for expandability.

Costing & Staffing

Cost table

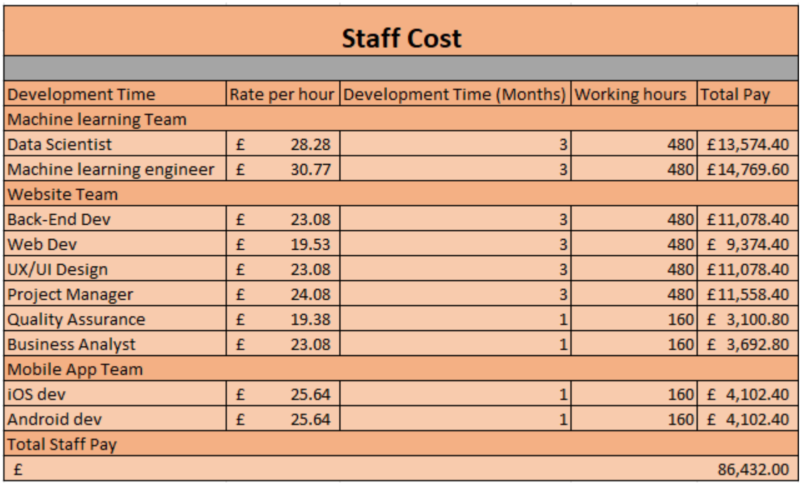


Project Cost sheet

Costs	Year 1	Year 2	Year 3	Year 4	Year 5	Total
Non-Recurring Costs						
Hardware costs						
Server	£ 1,359.98	£ -	£ -	£ -	£ -	£ 1,359.98
Server Rack	£ 396.00	£ -	£ -	£ -	£ -	£ 396.00
Harddrives	£ 199.47	£ -	£ -	£ -	£ -	£ 199.47
Boot SSD	£ 140.27	£ -	£ -	£ -	£ -	£ 140.27
Additonal Memory	£ 62.44	£ -	£ -	£ -	£ -	£ 62.44
Dual port Ethernet Adapter	£ 118.49	£ -	£ -	£ -	£ -	£ 118.49
1U 24 port Network switch	£ 77.74	£ -	£ -	£ -	£ -	£ 77.74
Software						
Ubuntu Server	£ -	£ -	£ -	£ -	£ -	£ -
MySQL	£ 3,745.50	£ -	£ -	£ -	£ -	£ 3,745.50
Staff Costs						
Machine Learning Team	£28,344.00	£ -	£ -	£ -	£ -	£ 28,344.00
Website Team	£49,883.20	£ -	£ -	£ -	£ -	£ 49,883.20
Mobile App Team	£ 8,204.80	£ -	£ -	£ -	£ -	£ 8,204.80
Total Non-recurring costs						
	£92,531.89	£ -	£ -	£ -	£ -	£ 92,531.89
Recurring Costs						
Software costs						
MySQL	£ -	£ 3,745.50	£ 3,745.50	£ 3,745.50	£ 3,745.50	£ 14,982.00
Ubuntu Advantage	£ 1,123.65	£ 1,123.65	£ 1,123.65	£ 1,123.65	£ 1,123.65	£ 5,618.25
Total reccuring costs						
	£ 1,123.65	£ 4,869.15	£ 4,869.15	£ 4,869.15	£ 4,869.15	£ 20,600.25
Overall project cost						
£						113,132.14

Staff Cost

Average salaries were used from uk.talent.com



Staff Cost

Development Time	Rate per hour	Development Time (Months)	Working hours	Total Pay
Machine learning Team				
Data Scientist	£ 28.28	3	480	£13,574.40
Machine learning engineer	£ 30.77	3	480	£14,769.60
Website Team				
Back-End Dev	£ 23.08	3	480	£11,078.40
Web Dev	£ 19.53	3	480	£ 9,374.40
UX/UI Design	£ 23.08	3	480	£11,078.40
Project Manager	£ 24.08	3	480	£11,558.40
Quality Assurance	£ 19.38	1	160	£ 3,100.80
Business Analyst	£ 23.08	1	160	£ 3,692.80
Mobile App Team				
iOS dev	£ 25.64	1	160	£ 4,102.40
Android dev	£ 25.64	1	160	£ 4,102.40
Total Staff Pay				
£				86,432.00

Legal Considerations

Data Protection Act

The user will create accounts that will be stored in a database. The account will contain personally identifiable information such as names and bank details which means only system administrators with security clearances in place may view that database. The code written to interact with it should not explicitly display that information in different components of the database or website and those identifiers must not be introduced into training the machine learning model.

Copyright and Licence

The Dataset is published under the Creative Commons universal license (Creative Commons, 2021) which allows it to be used with no copyright infringement on the author. The license allows the dataset to be modified, distributed, and used commercially without needing any permissions.

References

Algorithmia, 2020. *2020 state of enterprise machine learning.* [Online]
Available at: https://info.algorithmia.com/hubfs/2019/Whitepapers/The-State-of-Enterprise-ML-2020/Algorithmia_2020_State_of_Enterprise_ML.pdf

Creative Commons, 2021. *CC0 1.0 Universal (CC0 1.0).* [Online]
Available at: https://creativecommons.org/publicdomain/zero/1.0/

Didenko, O., 2021. *How Long Does It Take to Develop a Web.* [Online]
Available at: https://gbksoft.com/blog/how-long-does-it-take-to-develop-a-web-app/

IBM, 2020. *Mobile Application Development.* [Online]
Available at: https://www.ibm.com/cloud/learn/mobile-application-development-explained

Patel, A., 2021. *Customer Personality Analysis.* [Online]
Available at: https://www.kaggle.com/imakash3011/customer-personality-analysis/activity

Subhash, D., 2020. *What is The Ideal Software Development Team Size.* [Online]
Available at: https://www.it4nextgen.com/what-is-the-ideal-software-development-team-size/

Software References

Ubuntu server - https://ubuntu.com/download/server

Ubuntu Advantage - https://ubuntu.com/support

Hardware References

Server Rack - https://www.rackcabinets.co.uk/collections/value-server-racks/products/12u-value-server-rack-600-x-900

Novatech Server - https://www.novatech.co.uk/newmodserver.html?s=SR-0246

Network Switch - https://www.dcdi.co.uk/tp-link-rackmount-switches-pure-gigabit-unmanaged