Assignment 2

Data Analytics Challenge

Cohort 2 Group 1

Joyce Njeri, Christine Wasike, Alice Fatmata
Ayo Oluwapamilerin, Gilbert Sibomana

Department: Computer Science
University: African Leadership University
Country: Rwanda
Date: March 16, 2021

Contents

1 Introduction

1.1 Purpose

This document is Group 1 Cohort 2's bid to Rwanda Revenue Authority (RRA), which explains the implementation of our prototype, the environment we need to scale the project to production, our team's budget, timelines, risk, and recommendations for the successful implementation of the project.

1.2 Document Convention

This document follows the Latex format.

2 Stakeholders

2.1 Client

- Name: Rwanda Revenue Authority (RRA)
- Address: Kimihurura, Kigali, Rwanda
- Phone Number: (250) 788 185 500
- Email Address: info@rra.gov.rw

2.2 Contractor

- Name: 1C2(Cohort 2 Group 1)
- Address: KG 126 St, Kigali, Rwanda

- Phone Number: (250) 784 650 219
- Email Address: info@alueducation.com

3 Prototype

3.1 Background Information

The aviation industry is substantial for every country's revenue system as it is the gateway to getting people in and out of the country. One of the biggest challenges in the aviation industry is that it is costly to maintain a good standard due to financial hurdles and legal complications [2].

In 2015 Virgin America, a US-based airline, experienced customer loss and profit loss when the government issued a license tax on special fuel(airplane fuel). This made special fuel expensive for Virgin America, who decided not to increase ticket fares for customers but to cut costs. Unknowing to them, their cost-cutting affected their services, and they ended up losing customers [5].

Not only did they lose customers, but they also had their image tarnished as customers took it to Twitter to air their displeasure [2].

3.2 Implementation

3.2.1 Data Collection

We researched using the prompt to know the kind of data that is needed to solve the problem. After our research, we downloaded data that we found relevant to solving the problem. We used Pandas' to import the data set to Jupyter Notebook, our Artificial Intelligence environment. After importing it to Jupyter Notebook we used exploratory data analysis (EDA) to analyze, investigate and summarize the main characteristics of the data set [1].

3.2.2 Data Preparation

There are different ways to handle missing values: from dropping them, to filling them with zero, the mean, mode, or the median [2]. Slicing is the process of selecting specific rows and columns of data based on some criteria. We used it in data preparation to highlight the unique rows and columns we wanted to work with [2].

3.2.3 Data Transformation

Due to long column names, we used the Pandas' rename method to get short and simplified titles.

3.2.4 Data Integration

We integrated the sentiment data set with the financial data set in order to draw correlations and inferences.

3.2.5 Data Visualization

We have used Matplotlib [4] and Seaborn libraries in Python to visualize the data and to find a trend in the tax revenues collected by the United States government [1].

4 Environment

4.1 Hardware

- Computer or a Laptop

4.2 Software

- Jupyter Notebook
- Python libraries
- Operating System of your choice

5 Administration

5.1 Budget

Below is the estimated manpower budget, assuming that software is open-sourced and computer hardware is available to the client.

Expenses	Cost(Per month / USD)	Role Breakdown	Description
Data Analyst' Quotation	5000	Head Analyst Data Analyst	Money to pay the Data Analysts working on the product prototype back-end.
Software Development Quotation	9000	Head Developer Other Developers	Money allocated towards developing the product, i.e., Mobile Application, and the back-end system

Table 1: Budget Breakdown

5.2 Timelines

Below is the estimated project timeline.

Months	Tasks	Status
February	Data Preparation	Completed
March	Prototype Creation	Completed
April	Implementing And Testing Prototype	Not Started
May	Building Final Product	Not Started
June	Product Release	Not Started

Table 2: Project Timelines

5.3 Risk and Mitigation

Lack of adequate data may result in inaccurate findings. Having more data provides a more substantial basis for making fact-based decisions that can better inform company strategies.

To mitigate the risk of inadequate data, the institution must maintain a steady stream of data supply through customer collection.

6 Recommendations

We recommend Virgin America to increase their ticket prices and also restructure their target market to a customer demographic that can afford it.

As an alternative to increasing prices, we recommend that they should ensure proper allocation of funds to departments that interact directly with the customer base.

7 Conclusion

With the data set of our choice, we have been able to demonstrate how Exploratory Data Analysis (EDA) is the best way to understand and summarise the characteristics and behaviors of the data.

In summary:

- Plotting techniques help validate the hypothesis which is made about data.
- EDA helps us to understand which model will fit best for predictions about the data set.
- EDA also reduces our efforts at the time of machine learning model building

References

1 What is Exploratory Data Analysis?, Ibm.com, 2020. [Online]. Available: https://www.ibm.com/cloud/learn/exploratory-data-analysis. [Accessed: 04-Mar- 2021].

2 Part 4: Data Management and Analysis, Reporting and Disseminating Results. WHO STEPS Surveillance, 2017, pp. 4-1-1 to 4-4-1.

3 M. Wood, Python and Matplotlib Essentials for Scientists and Engineers, 3rd ed. Quebec: Morgan Claypool, 2015, p. 150pp.

4 ”What’s new in each version — seaborn 0.11.1 documentation”, Seaborn.pydata.org, 2021. [Online]. Available: https://seaborn.pydata.org/whatsnew.html. [Accessed: 05- Mar- 2021].

5 Z. Luvsandorj, ”Simple word cloud in Python”, Towards Data Science, 2020. [Online]. Available: https://towardsdatascience.com/simple-wordcloud-in-python-2ae54a9f58e5. [Accessed: 05- Mar -2021].