A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name: Cristhian Elson Pereira Macedo* |  |
| *Student Number: 2024104* |  |
| *Module Title: Strategic Thinking* |  |
| *Assessment Title: CA 3 Final Submission* |  |
| *Assessment Due Date: Sunday, 10 Nov 2024, 11:59 PM* |  |
| *Date of Submission: Sunday, 10 Nov 2024* |  |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.



**Amazon UK Products - Increasing Company Sales**

**Cristhian Elson Pereira Macedo**

**2024104**

Higher Diploma in Science in Data Analytics for Business

Strategic Thinking (STDA)

Lecturer: James Garza / Neil Doyle

CCT College Dublin

Dublin, Ireland

2024



**SUMMARY**

[**Table of Figures AND TABLES** iv](#_Toc182159668)

[**1.** **Introduction** 5](#_Toc182159669)

[**2.** **Objectives** 5](#_Toc182159670)

[**3.** **Problem Definition** 6](#_Toc182159671)

[**4.** **Scope** 6](#_Toc182159672)

[**5.** **Data Sources** 7](#_Toc182159673)

[**6.** **Ethical Considerations** 7](#_Toc182159674)

[**7.** **Data Understanding** 7](#_Toc182159675)

[**8.** **Data Preparation** 8](#_Toc182159676)

[**9.** **Modelling - Machine Learning Implementation** 10](#_Toc182159677)

[**10.** **EvaluationS - UPDATES OF THE PROJECT** 11](#_Toc182159678)

[**Conclusions** 18](#_Toc182159679)

[**timeline plotS** 22](#_Toc182159680)

[**References** 25](#_Toc182159681)

# **Table of Figures AND TABLES**

[Figure 1. Algorithm Comparison Train and Test Accuracy 13](#_Toc182159682)

[Figure 2. Algorithm Comparison Mean Squared Error Train and Test 14](#_Toc182159683)

[Figure 3. Algorithm Comparison: Kfold results 15](#_Toc182159684)

[Figure 4. Algorithm Comparison - Train and Test Error 16](#_Toc182159685)

[Figure 5. Algorithm Comparison - Train and Test R2 Score 17](#_Toc182159686)

[Figure 6. Comparison with 3 clusters between "K-Means Clustering", "K-Medoids Clustering" and "Fuzzy C-means Clustering”. 17](#_Toc182159687)

[Figure 7. Timeline Plot of Stage 1 - Project Plan / Business Understanding. 22](#_Toc182159688)

[Figure 8. Timeline Plot of Stage 2 - Data Understanding. 22](#_Toc182159689)

[Figure 9. Timeline Plot of Stage 3 - Data Preparation. 22](#_Toc182159690)

[Figure 10. Timeline Plot of Stage 4 - Machine Learning. 23](#_Toc182159691)

[Figure 11. Timeline Plot of Went past the deadline. 23](#_Toc182159692)

[Figure 12. Timeline Plot of Phase 5 - Updates/Analysis and Repeating previous phases. 23](#_Toc182159693)

[Figure 13. Timeline Plot of Phase 6 - Production of Report. 23](#_Toc182159694)

[Figure 14. Timeline Plot of Phase 7 - Presentation. 24](#_Toc182159695)

# **Introduction**

This Capstone Project will carry out a case study to understand and use a large dataset with over 2.2 million observations from one of the most popular online retailers around the world Amazon, the dataset used here was last updated in October of 2023, and it will be used to conduct the objects of this project using exploratory data analysis and machine learning to aim increasing company sales.

This large data set has been chosen to show the significance of managing a large amount of data and getting real results using a real-world problem. This study will be interesting and at the same time challenging, where one of the objectives is to develop possible solutions to train a model which creates an optimised title of the product making it interesting and more saleable than the large titles.

The dataset is composed of some objective features such as the title of the product, product rating, number of reviews, buy now the price of the product, whether the product had a status bestseller or not, number of products sold last month and the name of the category that the product belongs to.

Using the project management CRISP-DM methodology, “the most popular framework for executing data science projects.” (Data Science Process Alliance, 2021). Respecting the six phases: Business Understanding, Data Understanding, Data Preparation, Modelling and Evaluation.

The environment tool to be used in this project is Anaconda Navigator with Jupyter interactive computing notebook environment, using the programming language Python, with some technologies such as Pandas: For data analysis and manipulation, Numpy: For working with arrays, Seaborn: For statistical data visualisations, Matplotlib: For visualisation with Python, Scikit-Learn: For Machine Learning in Python, and Others: once needed it.

In addition, for spelling a free AI Writing Assistance Grammarly (Grammarly, 2009) is being used to help with English grammar while typing. The project is available on [this link](https://github.com/CCT-Dublin/capstone-project-feb-2024-ft-CristhianMacedo2024104) and the complete URL is available on References (Macedo, 2024).

# **Objectives**

This Capstone Project has a proposal objective to achieve and predict Increasing Company Sales, for this project, it will use the Amazon UK Products dataset by 2023 from the Kaggle website, some of the approaches to be covered are:

1. Analysing customer ratings with stars obtained from the products to show the best products by category and which niches are the easiest to make sales.
2. Figuring out trending products by category and their sales performance getting an idea of which products have the best sales performance.
3. Present possible recommendations to customers through the best-seller products with higher sales.
4. Train a model to predict and improve sales.
5. Attempt to train a generator of optimised titles with a short product name.

# **Problem Definition**

Nowadays due to the immense quantity of data generated by companies, it is getting harder to analyse and create an overview suitable for the best approaches to understand the business in the company and making the next steps to increase revenue. In addition, large text in the product “name” should be better described, easy to find during a search on the website, more interesting and more saleable than large titles.

# **Scope**

The first of two stages of Business Understanding in the project will be focused on understanding the environment and creating the documentation, the second stage will be to use the data, inspect the data set, describe, explore, verify data quality and so on, to become more familiar with the data frame.

The second stage will deal with data preparation, selecting data, cleaning, checking the types of data, missing and null values, irrelevant data, duplicates, type conversion, syntax errors, detecting outliers, standardise, scaling transformation, normalisation, dropping irrelevant columns, renaming the columns, using feature extraction (PCA or LDA), and so on when necessary.

The third stage is to create a Machine Learning model, using some selected modelling techniques such as regression, and random forest, generating a test design for getting into training, test and validation, and so on, the fourth is evaluating results, reviewing processes and determining the next steps of this capstone project.

After it is expected to deliver a machine learning model to predict with the best accuracy possible and more, any discoveries or different approaches used during the states of this project will be described in the document.

A timeline will be attached to this document once it begins and updated for each change made as GitHub commits with information such as task’s name, phases, progress percentages, start date and end date estimative for each task and so on, using a Simple Gantt Chart excel document (Vertex42, 2021), is expected at least one-two days to complete each task.

# **Data Sources**

The dataset mentioned before has been obtained from the website Kaggle, a community that allows you to “Discover a huge repository of community-published models, data & code for your next project” (Kaggle, 2019), this data is composed of 10 features and 2.2 millions of observations includes titles, number of reviews, ratings, prices, and sales data from October 2023.

This data set offers usability of 10.00, with an expected update frequency annually and this data set “Contains information from Amazon UK Products Dataset 2023 (2.2M Products) which is made available under the ODC Attribution License.”, this “[…] license agreement intended to allow users to freely share, modify, and use this Database subject only to the attribution requirements set out in Section 4”. (The Open Knowledge Foundation, n.d.)., the URIs from the data set and license agreement, both are available in the References section of this document using Harvard referencing.

# **Ethical Considerations**

Regarding ethical considerations associated with the data in this capstone project, there are no potential societal impacts, user privacy or any sensitive data to be used, all 10 original features seem within compliance without bias.

# **Data Understanding**

In this stage, it was imported the libraries pandas to manipulate the data and warnings to suppress any message, loaded the data set, inspected the data checked the Head and saw how the values were in each feature and the need to rename the feature’s name next got the shape with the size of 2222742 observations (Rows) and 10 features (Columns), got a summary with Info, obtained the result of numerical, Boolean and categorical features, checked the Describe with the statistic information about each numerical feature.

After using the isnull command obtained as result, no missing values between the features, also, plotted a graphic, checked duplicate observations, there were none, saw the nunique values of the features, understood some unique values of the features, and overviewing the data itself with a profile report.

After analyses, the feature Title contains large text and needs to be analysed, there exist some different scales comparing the min and max values between the features, which seems understandable due to the feature names Price, stars and reviews, and some categorical features that could be encoded.

# **Data Preparation**

After the first overview of the data set, it is time to do more analysis and prepare the data to follow the stages, after analyses, it opted not to delete any feature at the moment, as there are no tangible results for this yet, and if deleted there is no reason to test in techniques of feature Selection as UFS (Univariate Feature Selection) and RFE (Recursive Feature Elimination) or in Feature Extraction LDA (Linear Discriminant Analysis), PCA (Principal Component Analysis) or t-DSNE (t-Distributed Stochastic Neighbor Embedding).

Renamed the features, there are no syntax errors, missing values or duplicated observations to be handled at this stage, encoded the category feature in a new one, and replaced the values of the Best Seller feature, no type conversion to be dealt, checked the density, patterns and relationships of the features in the exploratory data analysis task, using a heatmap and a pair plot to visualise further, got all numeric features and checked the distribution of them in dist plot and kde plot, showing the statistical values Skew, Mean, Median and Standard Deviation, next tried the same using a bar plot, but if no success once the data is too large, calculations in a long period and also the results would not be properly represented.

Got a box plot of all numerical features in a graphic, next one by one and in a hist and box plot separately to understand more about the distribution and the existence of outliers, check a kde plot of the Category Encoded, and a scatter plot of Stars and Reviews features, next created analyses with the most popular products by category, first, it created a table of the top 10 categories grouped by stars and plotted in a bar plot graphic, got a table of the results and next doing analysis with Stars by Title, Stars by Reviews and Reviews by Stars.

It created another analysis in the graphic using a bar plot to find out which are the easiest niches to make sales, in this case, tested to get the 5 categories, after the results, the graphic presented that Sports & Outdoors, Skin Care, Fragrances, Makeup and Manicure & Pedicure Products are the best niches to make sales. Created a table to get the top 10 Titles with best-selling, and the results make sense where the products are from the Sports & Outdoors Category.

Next, it checked the top 10 Titles Bought in Las Month in a bar plot and another perspective in a table, and did the same as before, with Categories by Bought in Last Month. Also, a table of which best Bought Last Month's products was created to get an idea of which products had the best sales performance and plotted in a graphic.

Next, it created a copy of the original data set to test and handle outliers, to remove noise, variance or bias existing in the data, in this case, the features Reviews, PriceGBP, BestSeller and boughtLastMonth contain outliers as checked before, tested using IQR (Interquartile Range) and after calculations, it removed 737800 observations from the original 2222742, also used a boxplot to see the before and after, however, it opted not to remove any outliers, once these outliers could not be errors in the data, but just an unusual value.

In the next task, it tested scaling the features to get a normal distribution or as close as possible and get better performance in Machine Learning algorithms, kept the data after removing outliers to test and create a copy of new ones for new analyses, used MinMaxScaler, StandardScaler, Normalisation and RobustScaler techniques and always checking with a boxplot, it opted to keep RobustScaler once the algorithm is “[…] robust to outliers.” (scikit-learn developers, 2024a).

In the Feature Creation task, it was created new features such as a new column converted Pound to Euro, Using the feature Title, created: the count of numbers of words, the count of characters, the count of numbers, the count of Stop Words in English and the count of Special Characters. Also, there are no categorical features and time series to be handled.

Next on Feature Selection, it opted to Follow the line of reasoning of the Case Study "Advertisement click classification case study in Python" by "Farukh Hashmi" from "Thinking Neuron" (Hashmi, 2021) after analyses, it used a box plot and ANOVA test to check the correlations between the features Target “BestSeller” Categorical and the Numerical Features and next it used a bar chart and Chi-Square test as the target Categorical and the Categorical Features.

After analysis, it will reject the null hypothesis once there is evidence of a significant relationship to the following features: Numerical features: 'Stars', 'Reviews', 'PriceGBP', 'BestSeller', 'boughtLastMonth', 'CategoryEncoded', 'PriceEUR', 'TitleWordCount', 'TitleCharCount', 'TitleStopWords', 'TitleSpecialChar' and 'TitleNumCount' and Categorical features: 'Title', 'imgUrl', 'Category'.

Next, Feature Extraction opted to use LDA (Linear Discriminant Analysis), which returns the number of best features to be used, compared to PCA which needs to use the Elbow technique to analyse the best number of features that should be chosen. Encoded all categorical features to numeric, next plotted a heatmap to visualise it, next splitting the features in X and Y to train and test using “BestSeller” as a target variable.

After applying the values, the LDA algorithm returned one feature as a result, created a new data frame with this new feature and “BestSeller” as a target, applied in the following algorithms: Logistic Regression, Linear Discriminant Analysis, K Neighbors Classifier, Decision Tree Classifier, Gaussian NB, SVC, Random Forest Classifier and MLP Classifier, after test all algorithms got an excellent accuracy up to 99% and plotted in a box plot to difference scale between the Algorithm Comparison.

# **Modelling - Machine Learning Implementation**

After analysis with ANOVA test and Chi-Square test, it will reject the null hypothesis once there is evidence of a significant relationship to the features, which will be used in Machine Learning, but not considering the features: Asin, imgUrl and productURL, also there is no need to get the 'Category' feature, once had one feature encoded already, not getting the feature 'PriceGBP', instead to use GBP will be used the feature in EUR currency, also there is no need to get nominal features once it is going to be encoded, also, there is no need to create new numeric features using dummies.

After analyses until this step, it will be used the Machine Learning algorithms: Logistic Regression, Decision Tree, Random Forest and KNN - K Neighbors Classifier, to compare and Train a model to predict which product could be the next bestseller. It gets all features to “X” variable and Best Seller for “y” as predicted.

Logistic Regression scaled all features to the test, fit in the model and got good results in the Classification Report, Accuracy Score and Confusion Matrix, also, plotted a Confusion Matrix Display, a Confusion Matrix labelled from yellowbrick library, a Confusion Matrix and Classification Report in table form labelled, next a Class Prediction Error plot, a ROC graphic with the curves for Logistic Regression and a Class Balance graphic.

The Decision Tree did the same as before, it opted not to scale the features, fit in the model, got good results in the Classification Report, Accuracy Score and Confusion Matrix, also, plotted a Confusion Matrix Display and plotted a plot tree graphic.

The Random Forest did the same as the Decision Tree, got good results, plotted the trees in graphic and next used a Random Forest Regressor fit once again and plotted a Tree with just tree nodes.

The KNN - K Neighbors Classifier, the same except that, in this case, it was scalled all features to the test, got good results in the Classification Report, Accuracy Score and Confusion Matrix, also, plotted a Confusion Matrix Display and plotted a KNN Regression plot graphic with the 3 nearest neighbours.

Next, it used the Linear Regression Machine Learning algorithm, to predict sales using the Price in Euro feature and the Bought in Last Month feature as the predictor feature, fit the model got the results of Mean Squared Error and r square value, plotted in a scatter graphic also plotted a model prediction and created a Data Frame with Actual test values and Predicted values, also got the values of Mean Absolute Error, Mean Squared Error and Root Mean Squared Error.

# **EvaluationS - UPDATES OF THE PROJECT**

In this stage, it returned to phases 2, Data understanding, 3, Data Preparation and 4, Modeling, to create possible evaluations in the whole project. First, it was checked the objectives excluded some and updated the others, updated objectives 1, 2, 3, 5 and 6, deleted the first objective 4, the next 4 previously (5), and kept for future implementation objectives 6, 7 and 8. Analysed the whole document to validate it, changed colours to tons in Blue, corrected titles, checked objectives and questions, updated graphics, and analysed the efficiency of the document and data, tried to turn it better.

Instead of using 2 million observations, opted to use a sample to run it quickly, in this case, testing just with 5k observations when necessary, but is changeable if need to increase or decrease this number, tried using the chunk technique, but did not work well opted to create a fixed variable and attribute it a value, implemented Data efficiency, Model Comparison, Evaluation, Content-Based Recommender System, Machine Learning Unsupervised, updated Linear Regression algorithm, created a few analyses in different perspectives, created a Hyperparameter tuning and cross-validation, add the Local Interpretable Model-agnostic Explanations (LIME) for Machine Learning algorithms and tested SHapley Additive exPlanations (SHAP).

Updated the machine learning supervised algorithms: Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Random Forest Regressor and K Neighbors Classifier add LIME on all of them and SHAP in Random Forest Classifier, next created a Model Comparison of all getting the Accuracy and Mean Squared Error from Train and Test, also created another comparison with different machine learnings using K Fold with a value of 10 folds as a test.

Applied Evaluation (Hyperparameter tuning and cross-validation) in the "Decision Tree Classifier, Random Forest Classifier and Linear Regression" using Grid Search CV and Randomized Search CV, to see if anything can be improved in parameter matters, but in all cases, it just got worse results than before.

For Machine Learning - Unsupervised using PCA (Principal Component Analysis) to get two features from the data's signal and applied in K-Means Clustering, K-Medoids Clustering and Fuzzy C-Means Clustering with Comparison of Davies-Bouldin Index and Silhouette Score, also add a comparison of Clusters.

Updated Linear Regression algorithm question and perspective, tried Hyperparameter Tuning, but are not getting good results as analysed previously, the data are overfitting and imbalanced, so the results will not be good for predictions, tried to use Random Over Sampler to get the data balanced, also tried: “No oversampling, SMOTE, BorderlineSMOTE, SVM SMOTE, K-Means SMOTE” but the only method that works for this case is "Random Oversampling".

After applying the Random Over Sampler, it was possible to use a balanced data set with synthetic data, and using Linear Regression next improved a little bit the previous results.

Most of the questions are answered in topic 3.8.2 Analysis such as questions 1, 2 and 3, also created a few different analyses in this topic, question number 4 in topics 3.15.1 and 4.1, question number 5 in topic 4.9.

Below are a few results of Machine Learning algorithms. First, a table with an Accuracy score that shows all in 1.00 as predicted, in this analysis it was checked if a product could be a new Best Seller or not.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Logistic Regression** | **Decision Tree Classifier** | **Random Forest Classifier** | **Random Forest Regressor** | **K Neighbors Classifier** |
| Accuracy: 1.00 | Accuracy: 1.00 | Accuracy: 1.00 | Accuracy: 1.00 | Accuracy: 1.00 |

A graph of a graph of a train and test accuracy

Description automatically generated

Figure 1. Algorithm Comparison Train and Test Accuracy

Comparing the graphic above of Train and Test Accuracy, the best performer, could choose "Logistic Regression, Decision Tree Classifier, Random Forest Classifier, K Neighbors Classifier, Linear Discriminant Analysis and Ada Boost Classifier".

A graph with blue squares

Description automatically generated

Figure 2. Algorithm Comparison Mean Squared Error Train and Test

Comparing the graphic above of Train Error and Test Error, based on minor values, the best performs, could choose "LogisticRegression, KNeighborsClassifier and LinearRegression" they have a low train error and the test error in comparison to the others.

A graph of a graph

Description automatically generated with medium confidence

Figure 3. Algorithm Comparison: Kfold results

Above the K-Folds Cross Validation plotted in Boxplots is possible to analyse which algorithms had the best performance, based on minor average error, major accuracy value and boxplot with less variation, is possible to say that "Logistic Regression, KNeighborsClassifier, SVC and MLPClassifier" perform the best.

A graph with blue bars

Description automatically generated

Figure 4. Algorithm Comparison - Train and Test Error

Comparing Train Error and Test Error, based on minor values, the best perform, could choose "LogisticRegression, KNeighborsClassifier, SVC and MLPClassifier" they have a low train error and the test error in comparison to the others, unlike "DecisionTreeClassifier, RandomForestClassifier, RandomForestRegressor" having train error value too low in comparison of the test error too high, seem that the algorithm is overfitting.

A graph with blue squares

Description automatically generated

Figure 5. Algorithm Comparison - Train and Test R2 Score

Comparing Train R2 Score and Test R2 Score, based on values 1 or nearly 1, performs the best and below 0 shows lower variability of data, it could be chosen "RandomForestClassifier" as the best, also "DecisionTreeClassifier and RandomForestRegressor" others are basically zero.

A graph with different colored dots

Description automatically generated

Figure 6. Comparison with 3 clusters between "K-Means Clustering", "K-Medoids Clustering" and "Fuzzy C-means Clustering”.

In the graphic above it is possible to visualise that using "K-Means Clustering", "K-Medoids Clustering" and "Fuzzy C-means Clustering" are getting different clusters analysing the colours but all of them with 3 clusters or more are getting greats results.

# **Conclusions**

The first stage of this Capstone project started on March 14th and concluded on March 28th, according to the progress presented in “*Figure 7. Timeline Plot of Stage 1 - Project Plan / Business Understanding*”, it was focused on Project planning and Business Understanding. In the beginning, it was challenging to choose a good data set for this Study, where lots of data sets with different areas of expertise made it hard to pick just one, after searching and finally choosing one, it was started some reading and analysis about the dataset itself, also planned and started the Capstone Document limited to Word Count: 1,000 words, with the sections of Introduction, Objectives, Problem Definition, Scope, Data Sources, Ethical Considerations, Conclusions and References, for the first Assessment Task in the course.

Before starting the following stages of this Capstone project, it was corrected some topics in this document were after receiving feedback from the Lecturer, steps such as Objectives, Introduction, Problem Definition, Scope, Ethical Considerations and Timeline Plot were all properly updated and created if they did not exist before.

The second stage Data Understanding started on April 22nd and concluded on April 30th, according to the progress presented in “*Figure 8. Timeline Plot of Stage 2 - Data Understanding*”, it was focused on uploading necessary files and importing libraries, loading data, inspecting the data set, finding unique values and overviewing the data itself.

The third stage Data Preparation started on April 30th and concluded on March 18th, according to the progress presented in “*Figure 9. Timeline Plot of Stage 3 - Data Preparation*”, it was focused on Exploratory Data Analysis, with activities like dropping irrelevant columns, renaming columns, replacing values (syntax errors), handling/imputing missing values, handling duplicate rows, variable encoding, type conversion, handling outliers, feature scaling, feature creation, text data processing, time series features, feature selection (ANOVA test and Chi-Square Test), feature extraction and so on.

In this step, due to a large amount of information, it was challenging to manage it, it was spending considerable time learning some new things to apply in the project and at the same time having considerable patience time to wait for the analysis feedback, some steps such a present some graphs desired in the Exploratory Data Analysis task were not possible, due to the much information contained in the features, also using the LDA technique expend approximately an hour to complete the analyses and get the result between the algorithms tested, but even it was possible to complete the objectives 1, 2, 3 and 4.

The fourth stage Modelling (Machine Learning Implementation) started on March 18th and concluded on March 19th, according to the progress presented in “*Figure 10. Timeline Plot of Stage 3 - Machine Learning*”, it was focused on Machine Learning Algorithms, in this case, used Logistic Regression, Decision Tree, Random Forest, KNN - K Neighbors Classifier and Linear Regression, to conclude the objectives 5 and 6, which good results were obtained.

Due to unforeseen circumstances and not properly planning the time, the deadline for completing this project was exceeded, according to the progress presented in “*Figure 11. Timeline Plot of Went past the deadline*”, but with this, it was possible to finish all planned steps with more time, review all steps, learned and applied approaches, analysed better and write the ipynb and word document properly.

Some future recommendations of this Capstone Project are getting better results from objective 3, optimise the database for better analysis to expend less time waiting, evaluation step following CRISP-DM Phases of the Machine Learning algorithms in case it is needed, using GridSearchCV, hyperparameters, and so on, testing new approaches for analyses, new Machine Learning Algorithms approaches and the Deployment phase of the model as an API.

Furthermore, some limitations were faced as impediments to continuing the project and had to pause, objectives 8 and 9 for example, it is only possible to continue if it finished objective 7 first, due to the large text in the Title feature, a potential solution for future improvements should be dealt with this situation first decreasing text size, before to continue the previous recommendations. Finishing this document after adding a new Word count of ~2,000, Capstone Project is available on the GitHub website [on this link](https://github.com/CCT-Dublin/capstone-project-feb-2024-ft-CristhianMacedo2024104), also, all References from the ipynb and word document are available in both documents.

In the fifth stage Evaluation, started on October 28th and concluded on November 5th, according to the progress presented in “*Figure 12. Timeline Plot of Phase 5 - Updates/Analysis and Repeating previous phases*”, this stage was focused on the evaluation and turning the project better than before, with updates, new tests, analyses and implementations.

During this phase, it was important to identify strengths and weaknesses of the capstone project and apply improvements to change it better, challenges during the development of the stages were faced all the time, working with a large dataset is a big tournament, some techniques were tested to improve performance once the database in .csv file weighs 651.1 MB, to load it takes 4.23 seconds on the developer computer, using considerable CPU cores and allocating a lot of space in memory, such as 1.1 GB just at the beginning of the import, after analyses opted to use 2 M observations when necessary to understand and create basics analyses and when used in Machine Learning change to use a sample of it.

Created a few analyses to explore the data set and answered them all in the Exploratory Data Analysis topic, it updated the algorithms of Machine Learning Supervised (Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Random Forest Regressor and Neighbors Classifier) evaluation with Hyperparameter tuning and cross-validation, also applied LIME and SHAP to understand a little bit how the models are working.

Created another perspective using machine learning, in this time, with Unsupervised Machine Learning, using PCA (Principal Component Analysis) to create two features from the original data set getting the strong signal of it, next applied in K-Means Clustering, K-Medoids Clustering and Fuzzy C-Means Clustering, validating using the Davies-Bouldin Index and Silhouette Score, with a comparison of Clusters in the end.

Tried to apply a few techniques of Text Summarization such as "SUMY, BERT Extractive Summarization, T5 Abstractive Summarization, Gensim, TextTeaser and NLTK" to transform the title of products smaller to easily use it and visualise but did not get good results, so instead decided to create a manual temporarily solution to get the text title until finds a few symbols like: "|:–,". Also, created a simple Content-Based Recommender System using Cosine Similarity.

Also, it completed Phase 6 - Production of Report, which started on November 6th and concluded on November 8th, according to the progress presented in “*Figure 13. Timeline Plot of Phase 6 - Production of Report*”, and Phase 7 - Presentation, started on November 10th and concluded on November 11th, according to the progress presented in “*Figure 14. Timeline Plot of Phase 7 - Presentation*”.

The objectives until here were concluded, furthermore, a few recommendations, such as: removing emojis from the product name feature, creating a potential solution to train a generator of optimised titles to resolve this situation of large texts decreasing size, training a model to recommend the best products deeply and train a model to predict if a product could be bought with better accuracy and try to get around the problem of overfitting and imbalanced data.

Finishing this document after adding a new Word count of ~5,000, Capstone Project is available on the GitHub website [on this link](https://github.com/CCT-Dublin/capstone-project-feb-2024-ft-CristhianMacedo2024104), all References from the ipynb, word document and Poster are available in both documents. Finally, the video presentation is available [on this link](https://drive.google.com/drive/folders/1qzihv0SZoogog972l5_btjn5AGe27GLt?usp=sharing), also, the link is available in the References.

# **timeline plotS**

**A screenshot of a project report

Description automatically generated**

Figure 7. Timeline Plot of Stage 1 - Project Plan / Business Understanding.

*A screenshot of a computer

Description automatically generated*

Figure 8. Timeline Plot of Stage 2 - Data Understanding.

*A screenshot of a document

Description automatically generated*

Figure 9. Timeline Plot of Stage 3 - Data Preparation.

*A chart with a number of numbers

Description automatically generated with medium confidence*

Figure 10. Timeline Plot of Stage 4 - Machine Learning.

*A close-up of a graph

Description automatically generated*

Figure 11. Timeline Plot of Went past the deadline.

**A screenshot of a computer

Description automatically generated**

Figure 12. Timeline Plot of Phase 5 - Updates/Analysis and Repeating previous phases.

**A screenshot of a computer

Description automatically generated**

Figure 13. Timeline Plot of Phase 6 - Production of Report.

**A screenshot of a computer

Description automatically generated**

Figure 14. Timeline Plot of Phase 7 - Presentation.

# **References**

Albon, C. (2018). Machine learning with Python cookbook practical solutions from preprocessing to deep learning. Cambridge O’reilly.

aldemuro (2018). Comparing ML algorithms, train ACCURACY > 90%. [online] Kaggle.com. Available at: https://www.kaggle.com/code/aldemuro/comparing-ml-algorithms-train-accuracy-90#Introduction [Accessed Nov. 2024].

Analytics Vidhya (2019). Comprehensive Guide to Text Summarization using Deep Learning in Python. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2019/06/comprehensive-guide-text-summarization-using-deep-learning-python/ [Accessed 2024].

Asaniczka (2023). Amazon UK Products Dataset 2023 (2.2M Products). [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/asaniczka/amazon-uk-products-dataset-2023/data [Accessed 20 Mar. 2024].

Bordia, A. (2022). Handling Imbalanced Data by Oversampling with SMOTE and its Variants. [online] Analytics Vidhya. Available at: https://medium.com/analytics-vidhya/handling-imbalanced-data-by-oversampling-with-smote-and-its-variants-23a4bf188eaf [Accessed Nov. 2024].

CCT College Library (2023). Library: Academic Posters: Home. [online] Library.cct.ie. Available at: https://library.cct.ie/Posters/home [Accessed 6 Nov. 2024].

Cereniyim (2019). Distribution plot with mean and median. [online] GitHub Gist. Available at: https://gist.github.com/cereniyim/df1c8646411cc34eb9720a79400a200f [Accessed 11 May 2024].

Cristhian Macedo (2024a). CristhianMacedo\_SThinkingHDip\_CA3 Video. [online] CCT Google Drive. Available at: https://drive.google.com/drive/folders/1qzihv0SZoogog972l5\_btjn5AGe27GLt?usp=sharing [Accessed 10 Nov. 2024].

Data Science Process Alliance (2021). Evaluating CRISP-DM for Data Science. [online] Available at: https://www.datascience-pm.com/wp-content/uploads/2021/08/CRISP-DM-for-Data-Science.pdf.

Deb, D. (2020). Comparing the performance of different machine learning algorithms. [online] Dibyendu Deb. Available at: https://dibyendudeb.com/comparing-machine-learning-algorithms/ [Accessed Nov. 2024].

Domino Data Lab (2024). What is Feature Engineering? | Domino Data Science Dictionary. [online] domino.ai. Available at: https://domino.ai/data-science-dictionary/feature-engineering [Accessed 12 May 2024].

Gaurav, P. (2023). Step By Step Content-Based Recommendation System. [online] Medium. Available at: https://medium.com/@prateekgaurav/step-by-step-content-based-recommendation-system-823bbfd0541c [Accessed 2 Nov. 2024].

GeeksforGeeks (2018a). Get n-largest values from a particular column in Pandas DataFrame. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/get-n-largest-values-from-a-particular-column-in-pandas-dataframe/ [Accessed 23 May 2024].

GeeksforGeeks (2018b). Principal Component Analysis(PCA). [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/principal-component-analysis-pca/ [Accessed 18 May 2024].

GeeksforGeeks (2024). Hyperparameter Tuning in Linear Regression. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/hyperparameter-tuning-in-linear-regression/ [Accessed Nov. 2024].

Grammarly (2009). Grammarly. [online] Grammarly.com. Available at: https://app.grammarly.com/.

Hai Huang (2024). What are the best practices for feature selection in machine learning? [online] www.linkedin.com. Available at: https://www.linkedin.com/advice/3/what-best-practices-feature-selection-machine-u9crc [Accessed 12 May 2024].

Hashmi, F. (2021). Advertisement click classification case study in Python - Thinking Neuron. [online] Thinking Neuron. Available at: https://thinkingneuron.com/advertisement-click-classification-case-study-in-python/ [Accessed 16 May 2024].

Itamar Turner-Trauring (2019). Reducing Pandas memory usage #1: lossless compression. [online] Python⇒Speed. Available at: https://pythonspeed.com/articles/pandas-load-less-data/ [Accessed 30 Oct. 2024].

Kaggle (2019). Kaggle: Your Home for Data Science. [online] Kaggle.com. Available at: https://www.kaggle.com.

Keita, Z. (2023). Explainable AI - Understanding and Trusting Machine Learning Models. [online] Datacamp.com. Available at: https://www.datacamp.com/tutorial/explainable-ai-understanding-and-trusting-machine-learning-models [Accessed 4 Nov. 2024].

Macedo, C. (2024b). Capstone Project feb-2024 Ft CristhianMacedo2024104. [online] GitHub. Available at: https://github.com/CCT-Dublin/capstone-project-feb-2024-ft-CristhianMacedo2024104 [Accessed 2024].

Mahadevan, M. (2022). Step-by-Step Exploratory Data Analysis (EDA) using Python -. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2022/07/step-by-step-exploratory-data-analysis-eda-using-python/ [Accessed 11 May 2024].

Mihail Yanchev (2018). The most (time) efficient ways to import CSV data in Python. [online] Medium. Available at: https://medium.com/casual-inference/the-most-time-efficient-ways-to-import-csv-data-in-python-cc159b44063d [Accessed 31 Oct. 2024].

NumPy Developers (2024). Data types — NumPy v1.22 Manual. [online] numpy.org. Available at: https://numpy.org/doc/stable/user/basics.types.html [Accessed 31 Oct. 2024].

Płoński, P. (2020). How to visualize a single Decision Tree from the Random Forest in Scikit-Learn (Python)? [online] MLJAR. Available at: https://mljar.com/blog/visualize-tree-from-random-forest/ [Accessed 23 May 2024].

Sarowar Jahan Saurav (2023). 6 Useful Text Summarization Algorithm in Python - Sarowar Jahan Saurav - Medium. [online] Medium. Available at: https://medium.com/@sarowar.saurav10/6-useful-text-summarization-algorithm-in-python-dfc8a9d33074 [Accessed 2024].

scikit-learn developers (2024a). sklearn.preprocessing.RobustScaler — scikit-learn 0.24.2 documentation. [online] scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html [Accessed 11 May 2024].

scikit-learn developers (2024b). Univariate Feature Selection. [online] scikit-learn. Available at: https://scikit-learn.org/stable/auto\_examples/feature\_selection/plot\_feature\_selection.html [Accessed 12 May 2024].

Seaborn (2012). seaborn.kdeplot — seaborn 0.9.0 documentation. [online] Pydata.org. Available at: https://seaborn.pydata.org/generated/seaborn.kdeplot.html [Accessed 12 May 2024].

Shafi, A. (2023). Sklearn Random Forest Classifiers in Python Tutorial. [online] www.datacamp.com. Available at: https://www.datacamp.com/tutorial/random-forests-classifier-python [Accessed 23 May 2024].

SketchBubble.com (2024). Crisp DM. [online] www.sketchbubble.com. Available at: https://www.sketchbubble.com/en/presentation-crisp-dm.html [Accessed 6 Nov. 2024].

The Matplotlib Development Team (2024). matplotlib.pyplot.axvline — Matplotlib 3.8.4 documentation. [online] matplotlib.org. Available at: https://matplotlib.org/stable/api/\_as\_gen/matplotlib.pyplot.axvline.html#matplotlib.pyplot.axvline [Accessed 11 May 2024].

The Open Knowledge Foundation (n.d.). Open Data Commons Attribution License (ODC-By) v1.0 — Open Data Commons: legal tools for open data. [online] opendatacommons.org. Available at: https://opendatacommons.org/licenses/by/1-0/index.html.

The Pandas Development Team (2024a). pandas.DataFrame.memory\_usage — pandas 2.2.3 documentation. [online] Pydata.org. Available at: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.memory\_usage.html [Accessed 31 Oct. 2024].

The Pandas Development Team (2024b). pandas.DataFrame.reset\_index — pandas 1.4.2 documentation. [online] pandas.pydata.org. Available at: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.reset\_index.html [Accessed 11 May 2024].

The Pandas Development Team (2024c). pandas.DataFrame.T — pandas 2.1.3 documentation. [online] pandas.pydata.org. Available at: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.T.html [Accessed 11 May 2024].

Turing (2024). 5 Powerful Text Summarization Techniques in Python. [online] www.turing.com. Available at: https://www.turing.com/kb/5-powerful-text-summarization-techniques-in-python [Accessed 2024].

Vertex42, T. guide to excel in everything (2021). Simple Gantt Chart. [online] Vertex42.com. Available at: https://www.vertex42.com/ExcelTemplates/simple-gantt-chart.html?utm\_source=ms&utm\_medium=file&utm\_campaign=office&utm\_content=text#google\_vignette [Accessed 24 Mar. 2024].

Yashmeet Singh (2022). Boxplot With Separate Y-Axis for Each Column | Proclus Academy. [online] proclusacademy.com. Available at: https://proclusacademy.com/blog/quicktip/boxplot-separate-yaxis/ [Accessed 11 May 2024].