WARWICK BUSINESS SCHOOL
THE UNIVERSITY OF WARWICK

Masters Programmes: Group Assignment Cover Sheet

Student Numbers: Please list numbers of all group members	**5580065, 2155210, 5654007, 5663439, 5648495**
Module Code:	**IB9BW0**
Module Title:	**Analytics In Practice**
Submission Deadline:	**2nd December 2024, 12:00 pm**
Date Submitted:	**30th November 2024**
Word Count:	**1854**
Number of Pages:	**9**
Question Attempted: *(question number/title, or description of assignment)*	**Review Prediction: eCommerce Platform**
Have you used Artificial Intelligence (AI) in any part of this assignment?	**No**

Academic Integrity Declaration
We're part of an academic community at Warwick. Whether studying, teaching, or researching, we're all taking part in an expert conversation which must meet standards of academic integrity. When we all meet these standards, we can take pride in our own academic achievements, as individuals and as an academic community.

Academic integrity means committing to honesty in academic work, giving credit where we've used others' ideas and being proud of our own achievements.

In submitting my work, I confirm that:

- I have read the guidance on academic integrity provided in the Student Handbook and understand the University regulations in relation to Academic Integrity. I am aware of the potential consequences of Academic Misconduct.
- I declare that this work is being submitted on behalf of my group and is all our own, , except where I have stated otherwise.
- No substantial part(s) of the work submitted here has also been submitted by me in other credit bearing assessments courses of study (other than in certain cases of a resubmission of a piece of work), and I acknowledge that if this has been done this may lead to an appropriate sanction.
- Where a generative Artificial Intelligence such as ChatGPT has been used I confirm I have abided by both the University guidance and specific requirements as set out in the Student Handbook and the Assessment brief. I have clearly acknowledged the use of any generative Artificial Intelligence in my submission, my reasoning for using it and which generative AI (or AIs) I have used. Except where indicated the work is otherwise entirely my own.
- I understand that should this piece of work raise concerns requiring investigation in relation to any of points above, it is possible that other work I have submitted for assessment will be checked, even if marks (provisional or confirmed) have been published.
- Where a proof-reader, paid or unpaid was used, I confirm that the proof-reader was made aware of and has complied with the University's proofreading policy.

Upon electronic submission of your assessment you will be required to agree to the statements above

PROJECT REPORT

Table of Contents

I. Business Understanding

As part of our engagement with Nile, a leading eCommerce marketplace in Brazil, our objective is to develop a predictive model to help the company effectively identify customers likely to leave positive reviews. Nile has identified online reviews as a critical factor influencing consumer trust and purchasing decisions, ensuring its reputation remains strong.

This model will help Nile:

- Enhance customer engagement by focusing on those most likely to leave positive feedback.
- Optimise resource allocation by avoiding unnecessary communication with customers who are unlikely to leave positive reviews.
- Maintain a strong online reputation, which is vital for sustaining customer trust and driving sales.

Specifically, the model will:

- Predict the likelihood that a customer will leave a positively rated review, based on historical data.
- Focus on precision, aiming to minimise false positives, i.e., customers who are predicted to leave positive reviews but do not.
- Support Nile's marketing team by identifying customers who are the most promising candidates for review incentivisation, thereby optimising the company's resources.

To build an effective predictive model, Nile provided us access to eight datasets (along with a file for Product Category Translation) that include a wide range of information on customers, orders, and geolocation, among others.

We have to point out that the quality of the data will directly impact the model's accuracy. Therefore, it is essential to address any data gaps, inconsistencies, or noise before developing the model.

To address Nile's business needs, we will frame this as a predictive modelling task with a binary target variable: whether a customer would give a positive review (4-5 Stars) or a negative review (1-3 Stars). By treating this as a classification problem, we will assign each customer to one of the two categories based on the likelihood that their review will fall into that range. This will allow Nile to concentrate its efforts and focus on customers who are more

likely to provide good reviews. Given Nile's primary business objective is to maximise positive reviews, precision will be the key performance metric. In this case, precision will focus on minimising false positives, which refers to customers who are predicted to leave high ratings (4 or 5) but instead leave a lower rating (1, 2, or 3). We will focus on optimising the model to ensure that the predictions of positive reviews are as accurate as possible. For this reason, reducing false positives is particularly important because targeting customers who are unlikely to leave a positive review could result in wasted resources.

II. Data Understanding

From initial observations, we noticed that the data was well structured and segregated into simpler datasets. Through these datasets, we got a comprehensive understanding of all the data the client was accumulating regarding the customers and their purchases. The table below summarises the datasets.

Dataset Name	Dataset Description
olist_customers_dataset.csv	Customer identification number and address details
olist_sellers_dataset.csv	Seller identification number and address details
olist_orders_dataset.csv	Order-level information and key timestamps
olist_order_items_dataset.csv	Order, product, seller identification numbers, and order pricing
olist_order_reviews_dataset.csv	Review details and key timestamps
olist_products_dataset.csv	Product-level information
olist_order_payments_dataset.csv	Payment details for every order placed
olist_geolocation_dataset.csv	Customer address

After much inspection and consideration, we decided to use the customers, sellers, orders, order items, order reviews, and products dataset for modelling. The information in the geolocation dataset was also present in the other datasets so we decided to omit it. The

payments dataset had no useful information pertinent to solving the objective of this project so we decided not to use it.

While going through the order reviews dataset, we noticed a possible class imbalance among the review scores. The amount of 5,4 and 1 review scores were high, while the number of 2 and 3 review scores were surprisingly low in comparison. This observation will be addressed further in the report.

III. Data Preparation

In this phase, we joined all the data into one large dataset using the identification columns from each dataset. The date columns were converted to appropriate date-time values. We noticed some orders had an estimated date of delivery before the approval date and removed these observations but we would wish to discuss with the client to understand this discrepancy. We also removed outliers in the data such as those with early delivery of more than 80 days. Since we can't use *timedelta* variables for our model, we have used *totalseconds* to convert these *timedelta* variables into seconds, then divided the values by 86,400 to get the variables in the number of days.

We also checked the combined data for missing values. Out of 20,437 missing values, 11,187 were from the review score column. Since scores are the target variable, it was crucial to eliminate these gaps in data. Thus, we removed them from the dataset along with other minimal NA values we found in other columns. We noticed that some rows were exact duplicates and we made the decision to keep the first record and drop its duplicates.

We found the mean values for review scores for unique customer IDs and customer states which will be used in our modelling and removed any observations of unique customer IDs or customer states that only appear once to reduce data leakage and skewing of results.

When handling the training and test dataset split which we will discuss in our modelling steps later, we ensured that we predicted only the mean review scores for the training dataset and mapped it to the relevant unique customer IDs in the test dataset to reduce data leakage.

Next, we created a variable "Positive" that takes a value of 1 if the review score is 4 and above, and 0 otherwise. This helps reduce the class imbalance and will be our target variable during the modelling phase.

Finally, we have dropped columns we no longer require such as the different identification columns. All values of the final dataset have been converted to numerical values as our model doesn't work well with categorical data.

IV. Modelling

The model aims to predict if a customer would leave a positive review (Positive = 1). We utilised Random Forest and Gradient Boosted Decision Trees (GBDT) as they can generate and combine the results of multiple individual models to provide a more accurate prediction of a customer's expected rating. We then chose the model with better overall performance as our final model. We selected 7 features that we believe would have the largest impact on review scores. These features are largely categorised into three groups: Consumer, Seller, and Product characteristics, and are elaborated further in the table below.

Group	Characteristic
Customer	Mean score of customer's state
	Mean rating given by the customer
Seller	Number of days between estimated and actual day of delivery
	Number of days between approval of order and estimated day of delivery
	Number of days between approval of order and actual day of delivery
Product	Quantity of photos on the seller's page
	Price of the product

We first generated a model without any hyperparameters using a holdout method with a 20:80 test-train split, before proceeding to use a Random Search Function coupled with 5-Fold Cross-Validation to obtain the optimal hyperparameters for the models to increase the generalisability. Finally, we checked if the imbalance in the number of observations per class had any effect on our model's precision and took appropriate measures to handle it.

V. Evaluation

Our main metric for evaluation was macro precision, as we want to minimise the number of false positives, i.e. reducing the number of customers we think will give positive reviews but who give negative reviews instead. We proceeded to use Random Forest and GBDT models to model this problem. The macro precision for the unoptimised models when fitted to test data was 0.888 and 0.883 respectively, to three decimal places. With optimised hyperparameters, the precision values were 0.960 and 0.945, respectively, to three decimal places, when fitted to the training data. When predicting the test data, the precision scores were 0.888 and 0.886 respectively, to three decimal places. The classification report for each model's performance on the test data can be found in Figures 1 and 2 respectively.

```
Random Forest Best Params: {'max_depth': 21, 'max_features': 'log2', 'min_samples_split':
28, 'n_estimators': 55}
Model: Optimised Random Forest
Macro Precision: 0.887979518173424
Macro Recall: 0.8999823885926663
Macro F1-score: 0.8933282432311789


Classification Report:

              precision    recall  f1-score   support

           0       0.83      0.89      0.86       586
           1       0.94      0.91      0.92      1124

    accuracy                           0.90      1710
   macro avg       0.89      0.90      0.89      1710
weighted avg       0.90      0.90      0.90      1710
```

Figure 1

```
GBDT Best Params: {'criterion': 'squared_error', 'learning_rate': 0.15259491501025335,
'max_depth': 2, 'n_estimators': 30}
Model: Optimised GBDT
Macro Precision: 0.886279926335175
Macro Recall: 0.897422661630209
Macro F1-score: 0.891290527654164


Classification Report:

              precision    recall  f1-score   support

           0       0.83      0.89      0.86       586
           1       0.94      0.91      0.92      1124

    accuracy                           0.90      1710
   macro avg       0.89      0.90      0.89      1710
weighted avg       0.90      0.90      0.90      1710
```

Figure 2

Underfitting is unlikely as scores on the training data were high, and overfitting is also unlikely since the test scores did not differ by too much from training scores (Difference <0.1).

Finally, we ensured that the models were not affected by the apparent class imbalance between the review levels. In Figures 3 and 4 for Random Forest and GBDT respectively, we see that the model actively predicts for both classes. Additionally, the recall for both classes in either model never goes below 0.89. Hence, we safely concluded that class imbalance was not a major issue for either model.

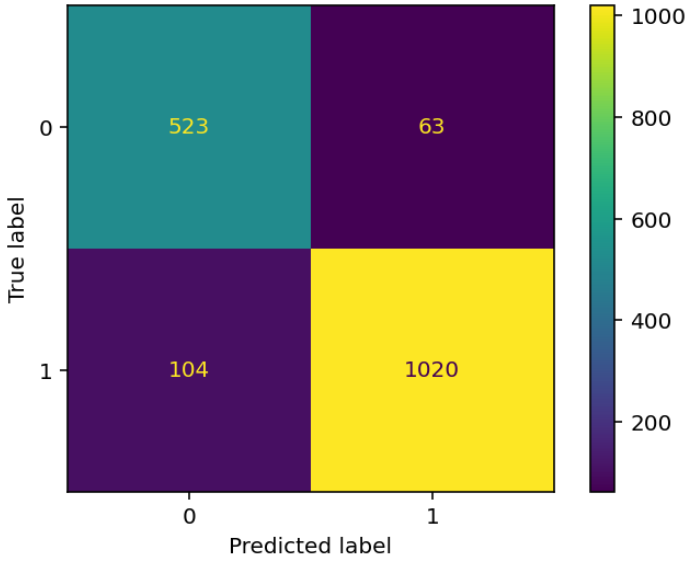

Figure 3

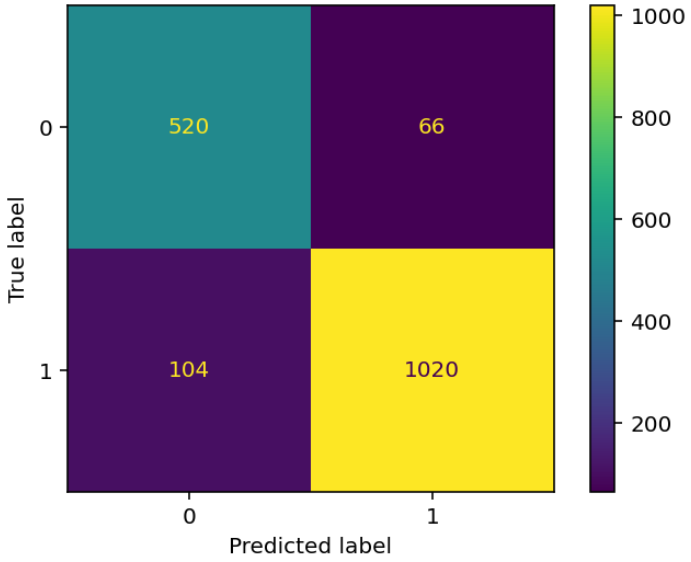

Figure 4

Ultimately, we opted for the optimised Random Forest model as both models have similar precision scores, but Random Forest has a better composition of feature importance. Figure 5 illustrates the feature importance chart of the GBDT model, where only the mean customer score is considered, which may indicate data leakage. Figure 6 illustrates the same but for the Random Forest, and while the mean customer score still takes a large proportion, it is much more well-balanced.

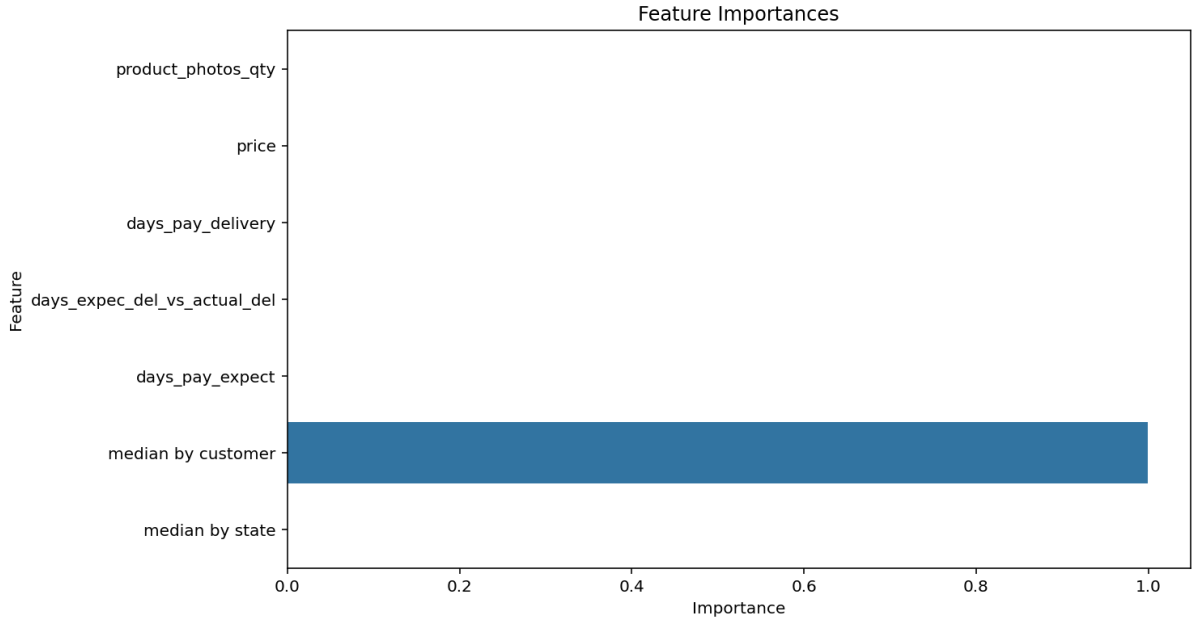

Figure 5

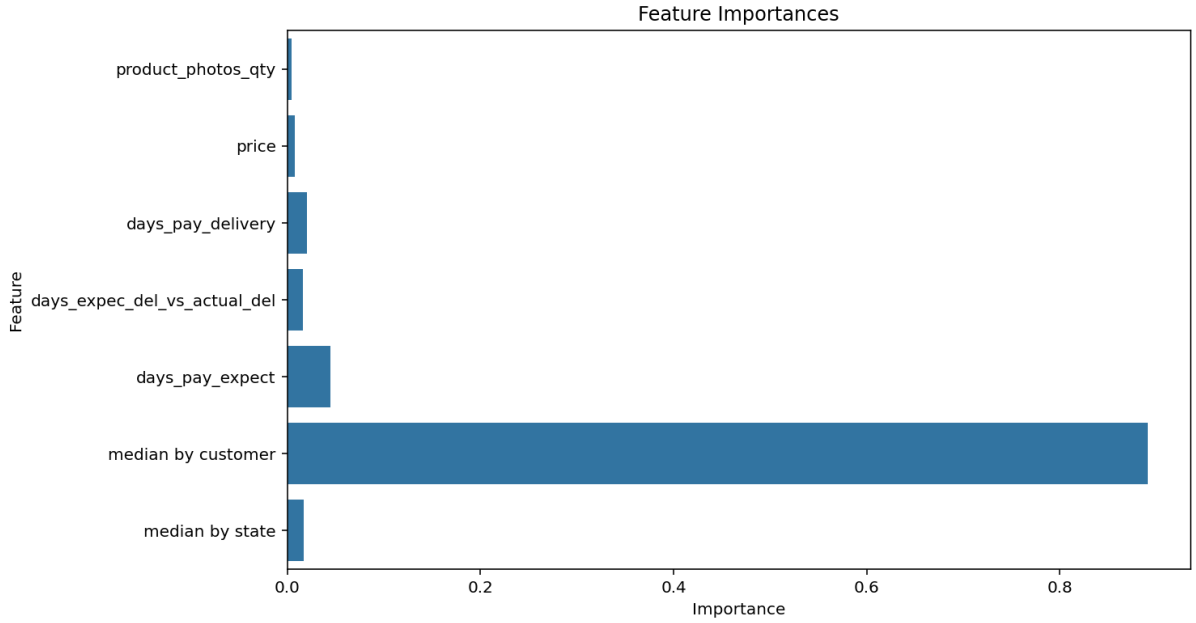

Figure 6

VI. Future Improvements

There are some future considerations and improvements for the model.

Firstly, finding out the reasons for the multitude of missing values, especially in the review_score column, and appropriately filling them instead of removing them will give the model more data to work with which may increase the precision of subsequent models. Additionally, behavioural data such as time spent on the page could also be beneficial for improving the accuracy of future models.

Secondly, this model is based on a smaller dataset of repeat customers to reduce data leakage when computing mean scores. This results in the model being effective when predicting reviews for repeat customers, but would likely be inefficient when predicting new customers. In the future, this model should be expanded with more features to boost precision scores and have more predictive power for new customers. It should also be trained on larger sets of data of recurring customers to ensure that it is generalisable on a larger scale.

VII. Deployment

After finalising the evaluation results, catering to the business's needs of prioritising precision and minimising false positives, the next step is integrating the optimised model into the production environment. The model can be deployed using an API to allow real-time predictions. Using fast API or flask, a REST API can be built around the model where POST and GET can accept a customer's data and return a review prediction. Once the model is deployed, monitoring should be continued to track the model's accuracy, recall, and F1 score to ensure it performs well.

Automation can tailor e-mails and incentives, thus engaging customers even better. This can be augmented with the implementation of the project and the campaigns for cross-selling and upselling. The model could also represent seasonal behaviour (e.g., with evaluation peaks during the festive seasons).