WARWICK BUSINESS SCHOOL
THE UNIVERSITY OF WARWICK

Masters Programmes: Group Assignment Cover Sheet

Student Numbers: Please list numbers of all group members	5672230, 5634661, 5593382, 5673642, 5622743, 5647195
Module Code:	IB9BW0
Module Title:	Analytics in Practice
Submission Deadline:	2nd December 2024
Date Submitted:	2nd December 2024
Word Count:	1925
Number of Pages:	16
Question Attempted: *(question number/title, or description of assignment)*	Develop and deploy a model to predict which customers will leave positive reviews
Have you used Artificial Intelligence (AI) in any part of this assignment?	OpenAI's ChatGPT was used to help with coding, fixing errors, and improving parts of the report to make them clear and accurate.

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

Table of Contents

1. Introduction

This project focuses on creating a prototype model using e-commerce data to predict which customers are most likely to leave positive reviews. The CRISP-DM (Cross-Industry Standard Process for Data Mining) method was used, which includes six main steps: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment (Provost and Fawcett, 2013). This report describes the steps taken to prepare the data, build the model, and test its performance.

2. Business Understanding

Nile, a large eCommerce platform, wants to predict which customers are likely to leave positive reviews to enhance brand reputation and optimise marketing efforts. The company aims to efficiently identify and encourage customers likely to leave positive reviews through targeted marketing. Current challenges include effectively managing incentives while minimising costs. By developing a predictive model, Nile seeks to optimise resource allocation, enhance customer engagement, and ensure consistent positive feedback.

3. Result and Discussion

3.1 Data Structure

The e-commerce dataset provided by Olist contains anonymised data on 100,000 Brazilian e-commerce orders (2016–2018) for comprehensive analysis (Sionek, 2011). Table 1 below are detailed descriptions of each dataset.

Table 1: Detailed information of dataset

Dataset	Information
Orders	Order details such as purchase, payment, delivery timestamps, and order status.
Order Items	Items purchased in each order
Product	Products sold by Olist.
Customer	Customers and their locations
Sellers	Sellers fulfilling orders for Olist, along with their locations
Payments	Orders payment options and amounts
Order Reviews	Customer reviews, including ratings and comments

In addition to the seven primary datasets, there are datasets translating product category names from Portuguese and geolocation data of Brazilian zip codes.

3.2 Data Exploration

The dataset shows 84,930 unique customers, 2,963 unique sellers, and 87,588 orders, with order counts peaking in November 2017. Most orders were delivered successfully, with 92.6% being delivered on time. Credit cards dominate payment methods, accounting for 75.9% of transactions. Furniture and Electronics emerge as the most popular product categories, and review scores clearly favour 4 or 5-star ratings, highlighting overall customer satisfaction. All this information will be useful for further analysis to predict which customers will give positive reviews (Appendix 1).

3.3 Data Preparation

The data preparation involved merging multiple datasets with inner joins to create a unified dataset, retaining only matching rows (see Appendix 2). This process combined customer, order, product, and seller data, excluding geolocation information due to its lack of relevance. This approach enabled a clear understanding of customer behaviour, product trends, and seller performance.

3.4 Feature Engineering

The dataset underwent several feature engineering steps to enhance its usability for predictive modelling:

1. Product Category Classification:
   Categorized the individual product categories into 9 broader categories: Furniture, Electronics, Fashion, Home & Garden, Entertainment, Beauty & Health, Food & Drinks, Books & Stationery, and Industry & Construction (Appendix 3).
2. Volume Calculation:
   A new column, product_volume, was created by multiplying product_length_cm, product_width_cm, and product_height_cm. This feature captures the overall size of a product, which could influence shipping costs or customer preferences.
3. Timestamp Conversion:
   Date columns were converted to a standardised datetime format. This allowed for easy calculation of time intervals using the format date/month/year hour: minutes. New features were created to allow a concise view on the delivery days
   a. Delivery Duration Days: Calculated as the difference between order_delivered_customer_date and order_purchase_timestamp. This captures the total time taken to deliver the product.
   b. Estimated Delivery Days: Difference between order_estimated_delivery_date and order_purchase_timestamp.
   c. Ontime Delivery: A binary feature indicating whether the delivery was on time (True if delivery_duration_days ≤ estimated_delivery_days).

d. Order Delay: Time difference between order_approved_at and order_purchase_timestamp, capturing delays in order approval.

4. Review Text Indicators:
   a. text_review: Created a binary indicator to denote whether a review contains a comment (yes or no).
   b. review_length: Captured the word count of the review comment to quantify review verbosity.
5. Customer Order Frequency: aggregated the total number of orders per customer by customer_id to provide customer purchasing behavior.
6. Review Categorisation and sampling:
   a. review: Transformed the numerical review score into categorical labels:
      - positive for scores above 3.
      - negative for scores of 3 or lower.

   To address class imbalance between positive and negative categories, SMOTE (Synthetic Minority Oversampling Technique) was used to oversample the negative review class for improved model training and performance, and reduced bias. This generates synthetic samples for the negative reviews, increasing their count to 60,000, while the positive class stays at around 70000.
7. Data encoding:

   Categorical variables in the dataset were transformed into a numerical format using one-hot encoding. This process created binary columns for each category within a variable, indicating the presence or absence of that category.
8. Scaling:

   Numeric feature values were scaled between 0 and 1, using Min-Max Scaling. This normalisation ensured that all variables contributed equally to the machine learning model, which could otherwise bias the model's learning process.

3.5 Data Cleaning

The data cleaning process involved identifying and removing columns that do not contribute to predictive modelling. The following steps were taken:

1. Dropping Non-Predictive Columns:

   A list of columns was defined for removal as these fields did not provide actionable insights or were irrelevant for prediction tasks. Specifically:
   a. Unique Identifiers: Columns like customer_id, order_id, and product_id were excluded as they are non-predictive unique identifiers.

b. Timestamp Data: Time-related fields were removed since relevant features (e.g., delivery_duration_days, ontime_delivery) were already engineered, making these raw columns redundant.
c. Text Fields: Columns like review_comment_message and review_comment_title were excluded due to the need for advanced preprocessing beyond the scope of this analysis.
d. Redundant/Non-Predictive Fields: Features like customer_zip_code_prefix and product_category_name were dropped for being irrelevant or superseded by more meaningful features.

2. Ensuring Data Completeness:
   After dropping the above columns, rows with missing values in remaining columns were removed along with duplicates, ensuring a clean dataset.

3.6 Modelling

The classification task involved training several advanced models, each chosen for its specific strengths in handling structured data. The selected models were:

1. Random Forest Classifier
2. Gradient Bossting Classifier
3. XGBoost Classifier
4. LightGBM Classifier
5. CatBoost Classifier

Each model's performance was evaluated based on a set of key metrics, designed to capture different facets of predictive accuracy and reliability (Precision, Recall, and F1-Score for Class 1). The table below summarises the models' performance on the training and testing data

Table 2: Model observations

Model	**Accuracy**		**Class 1**			**Observations**
	Training	**Testing**	**Precision**	**Recall**	**F1-Score**	
Gradient Boosting	0.8136	0.8124	0.79	0.90	0.84	Minimal overfitting, balancing training and test accuracies effectively.
XGBoost	0.8866	0.8727	0.84	0.94	0.89	Performed similarly, with slightly lower final scores due to marginally higher overfitting.
LightGBM	0.8761	0.8730	0.85	0.93	0.89	
CatBoost	0.8977	0.8847	0.86	0.95	0.90	Highest performance on the test set, achieving top scores

Model	Accuracy		Class 1			Observations
	Training	Testing	Precision	Recall	F1-Score	
						in recall and F1-score for Class 1, strong generalisation and suitability for positive sample identification.
Random Forest	0.9998	0.8673	0.85	0.92	0.88	Overfitting, with perfect accuracy on the training set but a noticeable decline in test accuracy

To improve model performance, a stacked ensemble approach was implemented by combining multiple boosting algorithms, aiming to achieve superior predictive accuracy. The performance of the stacked model is summarized below:

Table 3: Stacked Classifier

Model	Accuracy		Class 1			Observations
	Training	Testing	Precision	Recall	F1-Score	
Stacked Boosting	0.90	0.89	0.86	0.94	0.90	Performance similar to CatBoost

3.7 Assessment and Evaluation

The primary objective was to develop a predictive model to identify customers likely to leave positive reviews (Class 1), with a focus on precision, recall, and F1-scores for Class 1 while maintaining strong performance for Class 0. Gradient Boosting, XGBoost, LightGBM, CatBoost, and Random Forest were individually evaluated, and CatBoost was identified as the best-performing model due to its high recall and F1-score for Class 1, minimal overfitting, and strong generalization.

A stacked model combining boosting algorithms was then implemented and evaluated. While it demonstrated comparable performance to CatBoost, the improvement in metrics was marginal. Precision-recall curves and confusion matrices (as seen in Appendix 4) revealed that CatBoost achieved slightly better recall, while the stacked model had a marginal edge in precision. However, the increased complexity and computational requirements of the stacked model did not justify its use over CatBoost, and thus we picked CatBoost as our best performing model.

To refine CatBoost's performance further, Randomised Search Cross-Validation was applied to explore a range of hyperparameter values. The following parameters were optimized through RandomisedSearchCV to identify the most effective model configuration:

Table 4: Hyperparameter tuning

Parameter	Criteria	Optimal Value
learning_rate	Tested values between 0.01 and 0.3 for balancing convergence speed and stability.	0.19
depth	Values between 4 and 12, controlling the maximum depth of each tree and the model's complexity.	1
iterations	Examined between 500 and 2000 to optimise the number of boosting rounds.	1500
l2_leaf_reg	Adjusted as a regularisation term to mitigate overfitting.	12
border_count	Regulated the number of bins for numeric features, affecting the model's resolution.	64

The default CatBoost model was ultimately chosen for deployment because it demonstrated better generalization, stability, and efficiency compared to the hyperparameter-tuned model. While the tuned model achieved a marginally higher test accuracy (88.56% vs. 88.47%), it suffered from significant overfitting, achieving near-perfect accuracy on the training set (99.97%). This overfitting resulted from aggressive hyperparameters, including higher tree depth, faster learning rate, and excessive iterations, which compromised its ability to adapt to unseen data.

In contrast, the default CatBoost model maintained a balanced performance on both training and test sets, showcasing robustness and reliability. CatBoost's default settings are specifically optimized to handle diverse datasets without the risk of overfitting or unnecessary computational complexity. The minimal performance improvement from tuning did not justify the additional risks or computational costs, making the default model the more practical and reliable choice for real-world deployment.

3.8 Prototype Deployment

A prototype application was developed using Streamlit to deploy the CatBoost model for predicting customer review sentiment on the Nile eCommerce platform. The application enables users to input customer and order details, which are processed and encoded for prediction. The trained CatBoost model is used to predict whether a customer will leave a positive or negative review. Categorical variables are automatically encoded, and the

prediction result is displayed upon pressing the "Predict Review" button. This prototype provides a user-friendly interface for stakeholders to interact with the model and assess its performance in a practical context.

4. Future Work

After deploying the prototype, the next step will be developing a product development plan (Yang, 2023) to define project scope, align expectations and clarify roles and responsibilities of consultants and stakeholders. Consultants will conduct as-is analyses to define business requirements, develop and test the model. Nile will provide data support and review deliverables throughout the project. In addition, a change management plan should be developed in parallel with the model design phase to prepare stakeholders for changes that would occur following the implementation of the model.

Additional business requirements can be gathered during the as-is analysis phase to ensure that the model aligns with project objectives and is flexible enough to accommodate future business needs, such as leveraging predictive analytics and machine learning by combining insights from the company's historical data (Kaur, 2024), and integrating AI and cloud computing to enhance the customer shopping experience (Elmaaty et al., 2023).

5. Conclusion

Overall, we successfully developed a predictive model to help Nile identify customers likely to leave positive reviews. Among all the models tested, CatBoost emerged as the most effective, achieving the highest recall and balanced performance metrics for identifying positive reviews. The prototype is just a start, it confidently provides Nile with a valuable tool for targeted marketing and improved customer engagement thus we recommend this model if we are to move forward with the project.

References

Elmaaty, N. H. H. A. and Ibrahim, H. E. A. (2023) 'Integrating Artificial Intelligence and Cloud Computing in eCommerce Operational and Customer-Centric Advancements', *AI, IoT and the Fourth Industrial Revolution Review*, 13(9), pp. 18–28. Available at: https://scicadence.com/index.php/AI-IoT-REVIEW/article/view/17 (Accessed: 29 November 2024).

Kaur, R. (2024) 'Leveraging Machine Learning For Predictive Analytics In Ecommerce', *Educational Administration: Theory and Practice*, 30(6), pp. 536–543.

Provost, F. and Fawcett, T. (2013) *Data science for business: what you need to know about data mining and data-analytic thinking*. 1st edn. Sebastopol, CA: O'Reilly Media, p. 58.

Sionek, A. (2021) *Brazilian E-Commerce Public Dataset by Olist*. [online] Available at: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce (Accessed: 23 November 2024).

Yang, C. (2023) *Software Development Life Cycle*. [online] Available at: https://www.codecademy.com/resources/docs/general/software-development-life-cycle/prototype-model (Accessed: 25 November 2024).

Appendix 1. Data exploration

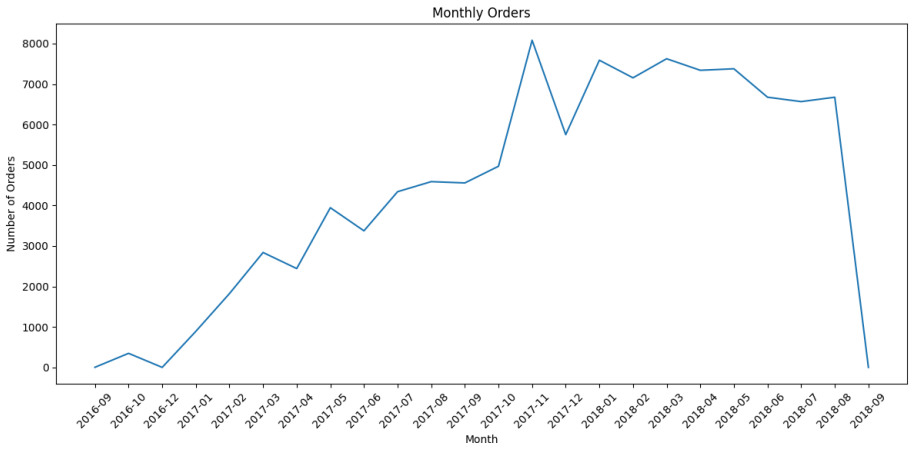

Figure 1: Monthly orders

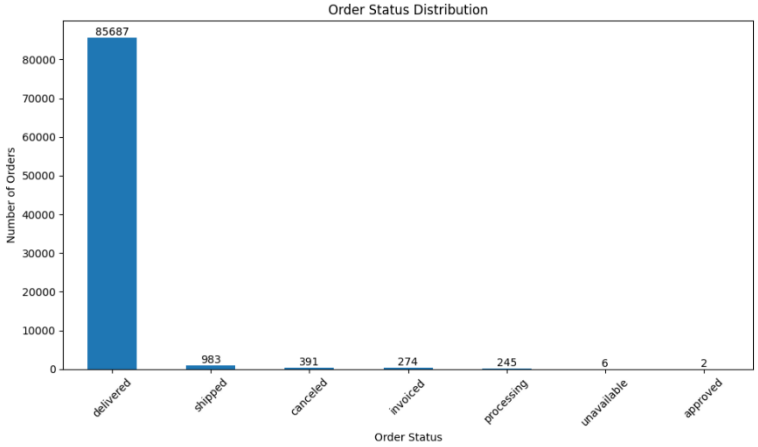

Figure 2: Order status distribution

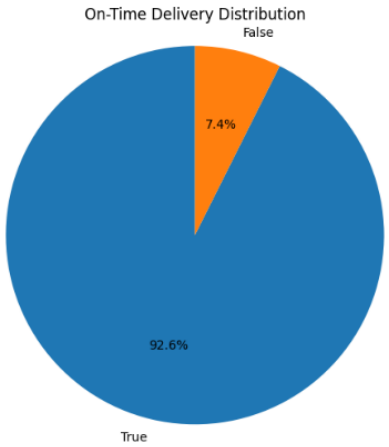

Figure 3: On-time delivery distribution

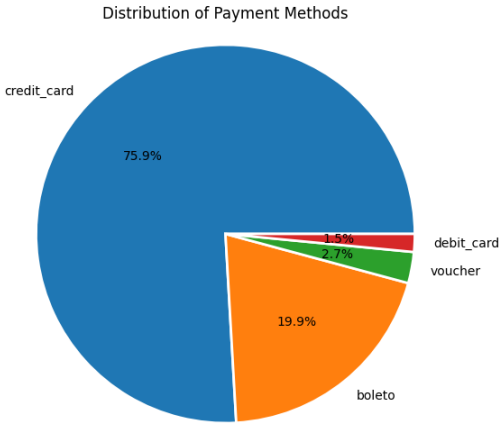

Figure 4: Payment method distribution

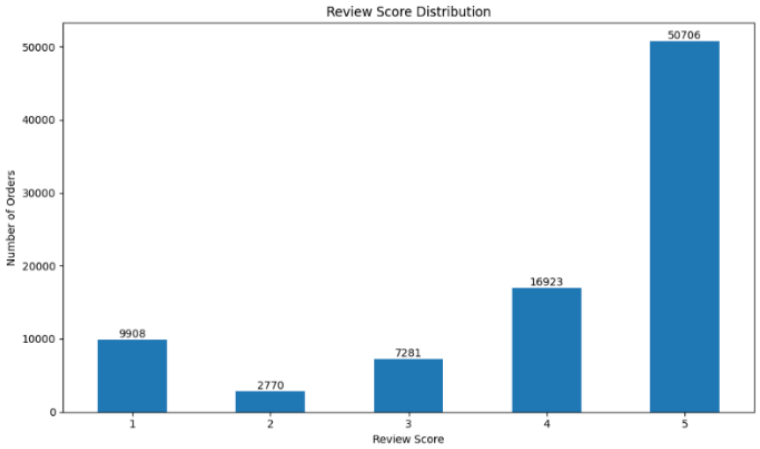

Figure 5: Review score distribution

Figure 6: Review score distribution by state

Appendix 2. Database Relation

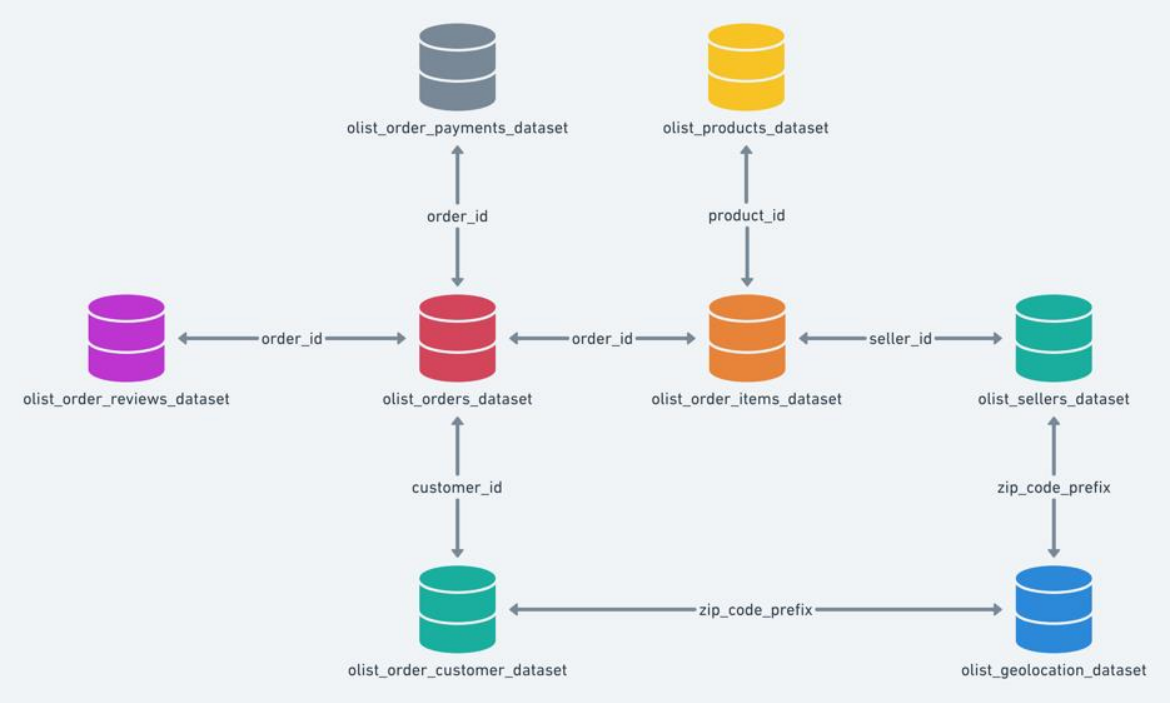

Figure 1: Database relation

Appendix 3. Product Category

Table 1: Product category classification

Category	Subcategories
Furniture	office_furniture, furniture_decor, furniture_living_room, kitchen_dining_laundry_garden_furniture, bed_bath_table, home_comfort, home_comfort_2, home_construction, garden_tools, furniture_bedroom, furniture_mattress_and_upholstery
Electronics	auto, computers_accessories, musical_instruments, consoles_games, watches_gifts, air_conditioning, telephony, electronics, fixed_telephony, tablets_printing_image, computers, small_appliances_home_oven_and_coffee, small_appliances, audio, signaling_and_security, security_and_services
Fashion	fashio_female_clothing, fashion_male_clothing, fashion_bags_accessories, fashion_shoes, fashion_sport, fashion_underwear_beach, fashion_childrens_clothes, baby, cool_stuff
Home & Garden	housewares, home_confort, home_appliances, home_appliances_2, flowers, costruction_tools_garden, garden_tools, construction_tools_lights, costruction_tools_tools, luggage_accessories, la_cuisine, pet_shop, market_place
Entertainment	sports_leisure, toys, cds_dvds_musicals, music, dvds_blu_ray, cine_photo, party_supplies, christmas_supplies, arts_and_craftmanship, art
Beauty & Health	health_beauty, perfumery, diapers_and_hygiene
Food & Drinks	food_drink, drinks, food
Books & Stationery	books_general_interest, books_technical, books_imported, stationery
Industry & Construction	construction_tools_construction, construction_tools_safety, industry_commerce_and_business, agro_industry_and_commerce

Appendix 4. Model Analyses

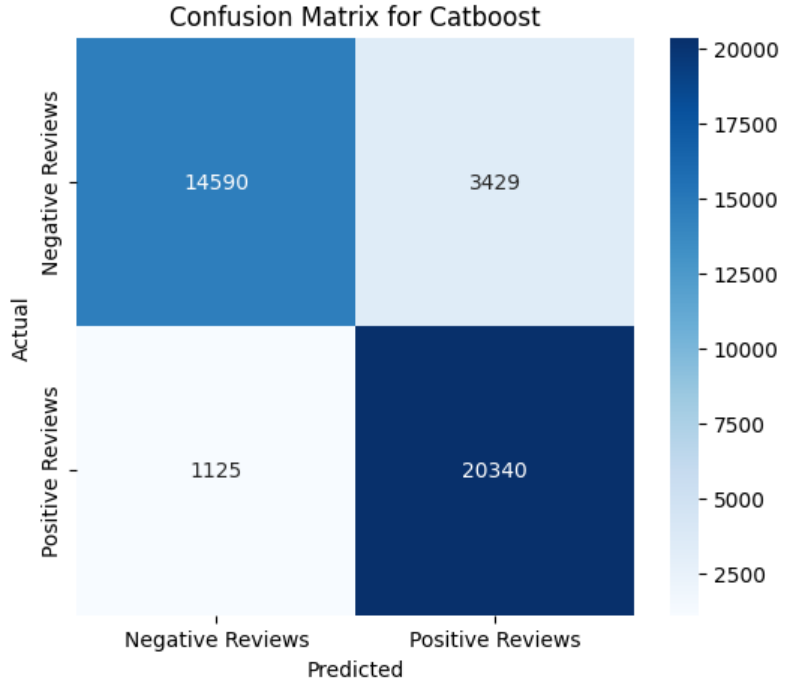

Figure 1: Confusion Matrix for Catboost

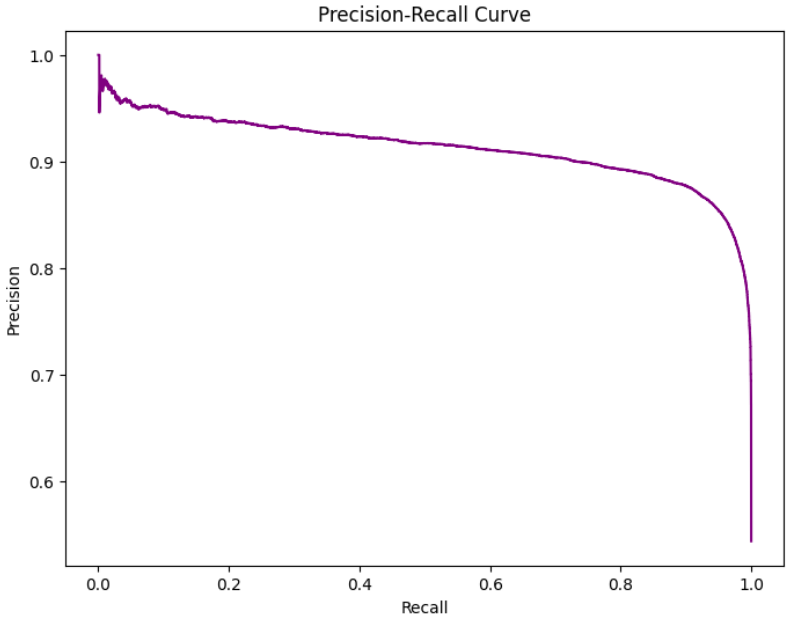

Figure 2: Precision-Recall Curve for Catboost

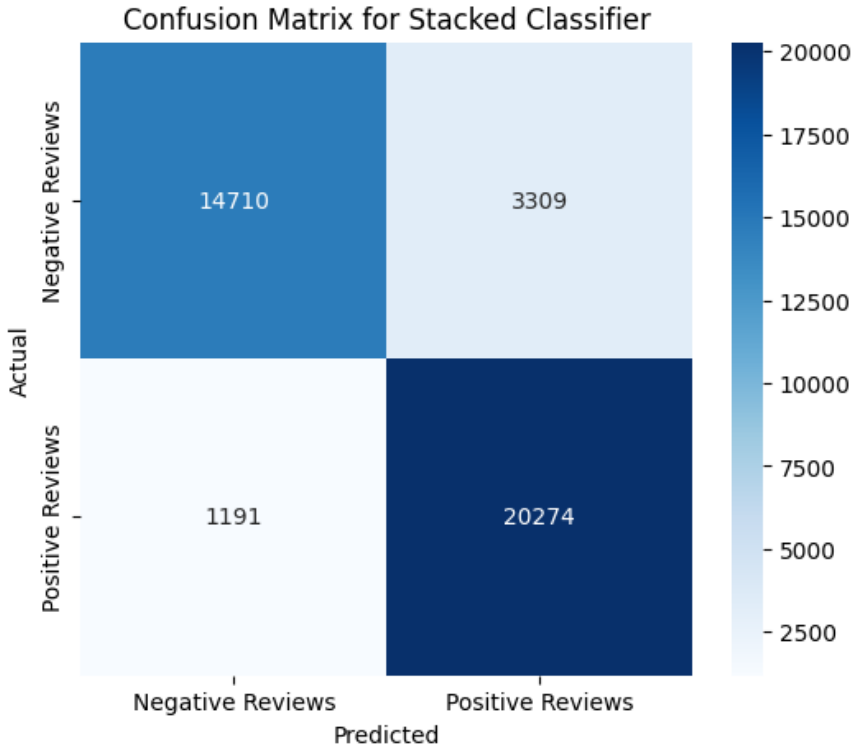

Figure 3: Confusion Matrix for Stacked Model

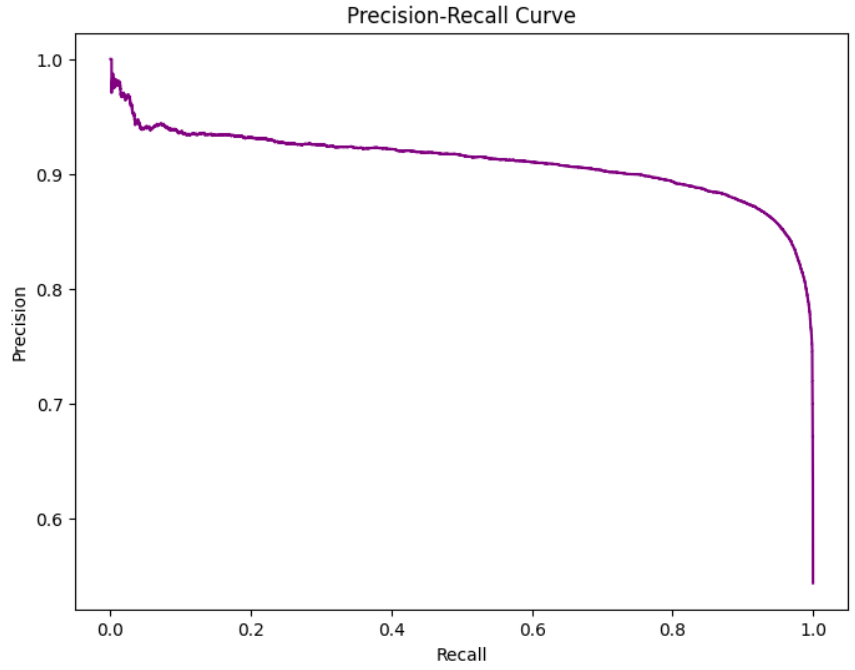

Figure 4: Precision-Recall Curve for Stacked Classifier