Data Science Intern

PREDICTING CUSTOMER SATISFACTION

submitted by

Sanjil K C – (SISPLINT: 121020345)

in partial fulfilment for the award

of

Data Science Intern
Shiash Info Solutions Private Limited

Shia#sh
Focus on Deliverables

March 2024

ACKNOWLEDGEMENT

In successfully completing this project, many people have helped me. I would like to thank all those who are related to this project.

I would like to thank **Dwaraka Reddy**, Professor, Shiash Info Solutions Private Limited, Chennai. He gave me this opportunity to work on this project and their patience, motivation, enthusiasm, and immense knowledge insisted me to put forward my fullest effort. His guidance helped me in all the time of research and implementation of the project. I got to learn a lot from this project about how to search for the best datasets for the projects, how to make your project more effective, which will be very helpful in solving some real-world problems and many other things.

At last, I would like to extend my heartfelt thanks to my parents because without their help this project would not have been successful. Finally, I would like to thank my dear friends and my fellow students for many helpful discussions and good ideas along the way who have been with me all the time.

CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATION

Abbreviation	Expansion
EDA	Exploratory Data Analysis
API	Application Programming Interface
ML	Machine Learning
UI	User Interface
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
CV	Cross-Validation
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
RF	Random Forest

1 INTRODUCTION

In the vast and bustling realm of e-commerce, where transactions flow ceaselessly and consumer preferences shift like the tides, Olist stands as a stalwart sentinel of innovation and progress. Nestled within the vibrant tapestry of Brazil's marketplace, Olist serves as a linchpin, connecting merchants and consumers in a symphony of commerce and connectivity.

Amidst the hustle and bustle of online transactions, Olist recognizes a fundamental truth: the bedrock of success in e-commerce lies in the realm of customer satisfaction. For Olist, every interaction, every purchase, and every review is not merely a transaction but an opportunity—an opportunity to delight, to exceed expectations, and to forge enduring connections.

Yet, in the ever-evolving landscape of digital commerce, the quest for customer satisfaction is no easy feat. It requires more than just operational efficiency and a robust infrastructure—it demands foresight, insight, and a keen understanding of the intricate dance between consumer desires and marketplace dynamics.

In this narrative of innovation and adaptation, Olist sets forth on a journey—a journey fueled by data, propelled by curiosity, and guided by a singular mission: to anticipate, understand, and surpass customer expectations at every turn.

With a vast trove of data spanning years of transactions and interactions, Olist embarks on a voyage of discovery—a quest to unlock the secrets hidden within the numbers, to discern patterns amidst the noise, and to glean insights that transcend the mundane.

In the corridors of data, Olist seeks not just information but illumination—insights that illuminate the path forward, guiding strategic decisions and shaping the trajectory of the business. Through rigorous analysis and sophisticated modeling techniques, Olist endeavors to peer into the future, foreseeing customer sentiments before they are voiced and crafting tailored solutions to meet their needs.

In this narrative arc of exploration and innovation, Olist emerges as a beacon—a beacon of excellence, of ingenuity, and of unwavering commitment to customer satisfaction. With each data point and each analysis, Olist charts a course toward a future where customer delight is not just a goal but a defining characteristic of the e-commerce experience.

2 EXISTING SYSTEM

In the dynamic landscape of e-commerce, where customer preferences and market trends shift with remarkable speed, the existing system grapples with the formidable challenge of staying ahead of the curve. At the heart of this system lies a robust framework for collecting and analyzing customer feedback, a cornerstone of Olist's commitment to excellence.

1. Post-Transaction Feedback Mechanisms: At the core of the existing system are post-transaction feedback mechanisms designed to capture customer sentiments in the aftermath of a purchase. These mechanisms, ranging from email surveys to in-app rating prompts, serve as invaluable touchpoints for gathering insights into customer satisfaction levels.

2. Manual Analysis and Interpretation: Following the collection of feedback, Olist's dedicated team of analysts embarks on a journey of exploration, meticulously dissecting each review, rating, and comment. Through manual analysis, patterns emerge, trends crystallize, and actionable insights come to light. However, this process, while thorough, is labor-intensive and time-consuming, limiting the agility of response mechanisms.

3. Iterative Improvement and Adaptation: Armed with insights gleaned from feedback analysis, Olist embarks on a journey of iterative improvement, fine-tuning its operations, policies, and service offerings to better align with customer expectations. This cyclical process of refinement forms the backbone of Olist's customer-centric approach, ensuring continuous evolution and adaptation in response to changing market dynamics.

4. Customer Relationship Management (CRM) Systems: Underpinning the existing system are sophisticated CRM systems, meticulously designed to manage customer interactions, track feedback trends over time, and foster meaningful relationships with customers. These systems serve as invaluable repositories of customer data, enabling personalized communication and targeted outreach efforts.

While the existing system excels in its ability to capture and analyze customer feedback, it grapples with inherent limitations. Chief among these is its reactive nature, wherein insights are gleaned post-transaction, offering a retrospective view of customer satisfaction. Moreover, the reliance on manual analysis poses scalability challenges, hindering real-time responsiveness to emerging trends and issues.

As Olist charts its course toward a future of unparalleled customer satisfaction, there arises a pressing need to augment the existing system with predictive analytics and proactive measures. By harnessing the power of data-driven insights and predictive modeling, Olist aims to anticipate customer sentiments before they crystallize, paving the way for proactive interventions and transformative customer experiences.

3 PROPOSED SYSTEM

In response to the limitations of the existing system and the evolving demands of the e-commerce landscape, Olist proposes a transformative new approach—a forward-thinking system that leverages predictive analytics and proactive measures to anticipate and exceed customer expectations.

Key Components of the Proposed System:

1. **Predictive Analytics Framework:** At the heart of the proposed system lies a sophisticated predictive analytics framework, powered by machine learning algorithms and advanced statistical models. By analyzing historical transaction data, customer behaviors, and contextual variables, this framework forecasts future customer satisfaction levels with unprecedented accuracy.

2. **Real-time Sentiment Analysis:** As transactions unfold in real-time, Olist's system continuously monitors customer interactions and sentiments, leveraging natural language processing (NLP) techniques to analyze text-based feedback and sentiment. This real-time sentiment analysis enables Olist to identify emerging trends, address issues promptly, and proactively engage with customers to mitigate dissatisfaction.

3. **Personalized Recommendations:** Building upon insights gleaned from predictive analytics, Olist's system generates personalized recommendations and tailored offerings for individual customers. By understanding each customer's preferences, purchase history, and feedback, Olist delivers targeted recommendations that resonate with their unique needs and preferences, fostering deeper engagement and loyalty.

4. **Automated Response Mechanisms:** In tandem with real-time sentiment analysis, Olist's system employs automated response mechanisms to address customer concerns and resolve issues swiftly. By leveraging chatbots, automated emails, and self-service portals, Olist streamlines the customer support process, providing timely assistance and enhancing overall satisfaction.

5. **Continuous Learning and Optimization:** As transactions and interactions accrue, Olist's system undergoes continuous learning and optimization, refining its predictive models, algorithms, and response strategies over time. By iteratively adapting to changing market dynamics and customer preferences, Olist ensures that its system remains at the forefront of innovation and effectiveness.

Benefits of the Proposed System:

- **Proactive Customer Engagement:** By anticipating customer sentiments and needs in advance, Olist can proactively engage with customers, preemptively addressing concerns and fostering positive experiences.
- **Enhanced Operational Efficiency:** Through automation and real-time analytics, Olist's system streamlines processes, reduces response times, and optimizes resource allocation, enhancing operational efficiency and agility.
- **Improved Customer Satisfaction:** With personalized recommendations, timely responses, and proactive interventions, Olist's system cultivates a culture of customer-centricity, driving higher levels of satisfaction and loyalty.

- **Data-driven Decision Making:** By harnessing the power of data-driven insights, Olist gains deeper visibility into customer behaviors, market trends, and operational performance, enabling informed decision-making and strategic planning.

In essence, the proposed system represents a paradigm shift—a departure from reactive, hindsight-driven approaches toward proactive, foresight-driven strategies. By embracing predictive analytics and proactive engagement, Olist aims to redefine the e-commerce experience, delivering unparalleled value and satisfaction to customers while driving sustainable growth and success.

4 SOFTWARE REQUIREMENT

1. **Python:** Primary programming language for data analysis, machine learning, and model development.

2. **Integrated Development Environment (IDE):**

    a. **Jupyter Notebook:** Interactive environment for code development, data exploration, and visualization.

    b. **PyCharm or Visual Studio Code:** IDEs for code editing, debugging, and version control.

3. **Data Manipulation and Analysis Libraries:**

    a. **NumPy:** For numerical computations and array manipulation.
    b. **Pandas:** For data manipulation, preprocessing, and analysis.
    c. **Matplotlib and Seaborn:** For data visualization and exploratory data analysis (EDA).

4. **Machine Learning Libraries:**

    a. **Scikit-learn**: For building and evaluating machine learning models, including classifiers for binary classification tasks.
    b. **XGBoost or LightGBM:** For gradient boosting models, suitable for classification tasks.
    c. **TensorFlow or PyTorch:** For deep learning models if neural network architectures are explored.

5 HARDWARE REQUIREMENT

1. **Processor (CPU):**

   - A multi-core processor (e.g., Intel Core i5 or higher, AMD Ryzen 5 or higher) is recommended for efficient data processing and model training.
   - If you plan to train complex deep learning models, a CPU with higher clock speeds and more cores (e.g., Intel Core i7 or AMD Ryzen 7) may be beneficial.

2. **Memory (RAM):**

   - At least 8 GB of RAM is recommended for handling moderate-sized datasets and training machine learning models.
   - For larger datasets and more memory-intensive tasks, consider 16 GB or 32 GB of RAM for smoother performance.

3. **Storage (Hard Drive):**

   - A Solid-State Drive (SSD) is preferable over a Hard Disk Drive (HDD) for faster data access and processing.
   - Aim for at least 256 GB of SSD storage for storing datasets, code, and project files.
   - Additional external storage may be needed for larger datasets or backups.

4. **Internet Connectivity:**

   - Stable internet connectivity is essential for accessing online resources, downloading datasets, and deploying web-based applications.
   - Higher bandwidth connections may be necessary for cloud-based model training or deployment.

5. **Cooling System:**

   - Adequate cooling is important, especially if your system will be running intensive computations for prolonged periods.
   - Consider investing in additional cooling solutions, such as aftermarket CPU coolers or case fans, to maintain optimal operating temperatures.

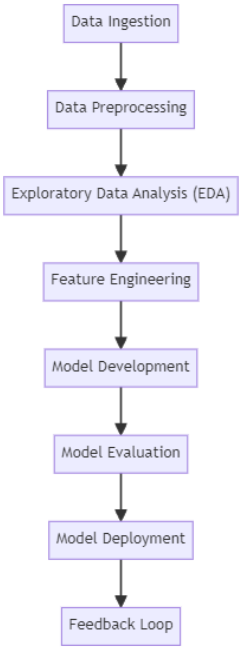
6. **Power Supply Unit (PSU):**

   - Ensure your power supply unit can handle the power requirements of your hardware components, especially if you're using a dedicated GPU.
   - Aim for a PSU with sufficient wattage and efficiency to power all components reliably.

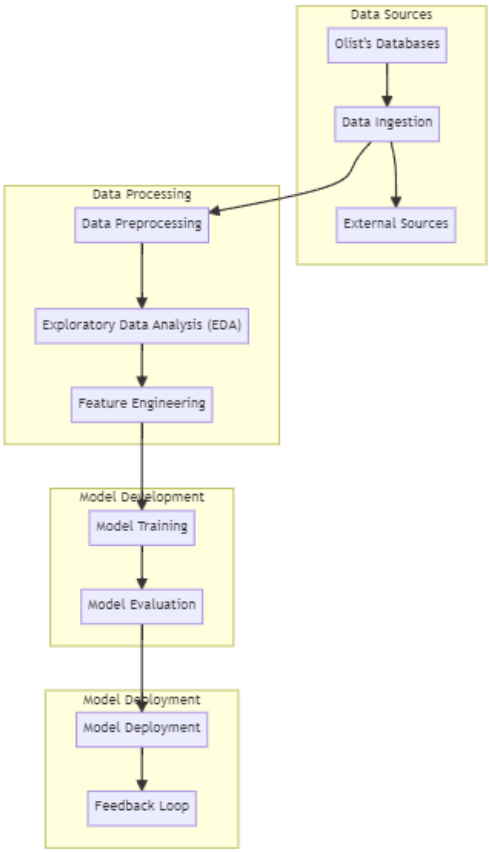
7. **Monitor:**

   - A high-resolution monitor with good color accuracy is essential for visualizing data, monitoring model training progress, and analyzing results effectively.
   - Consider a dual-monitor setup for increased productivity and workflow flexibility.

6 ARCHITECTURAL DIAGRAMS



7 DATA FLOW DIAGRAM



8 TABLE DESIGN

8.1 Customer Table

	customer_id	customer_unique_id	customer_zip_code_pr	customer_city	customer_state
2	06b8999e2fba1a1fbc8	861eff4711a542e4b93843	14409	franca	SP
3	18955e83d337fd6b2de	290c77bc529b7ac935b93a	9790	sao bernardo do cai	SP
4	4e7b3e00288586ebd08	060e732b5b29e8181a182	1151	sao paulo	SP
5	b2b6027bc5c5109e529	259dac757896d24d7702b9	8775	mogi das cruzes	SP
6	4f2d8ab171c80ec8364	345ecd01c38d18a9036ed9	13056	campinas	SP

8.2 Geolocation Table

1	geolocation_zi	geolocatio	geolocatio	geolocatio	geolocation_state
2	1037	-23.5456	-46.6393	sao paulo	SP
3	1046	-23.5461	-46.6448	sao paulo	SP
4	1046	-23.5461	-46.643	sao paulo	SP
5	1041	-23.5444	-46.6395	sao paulo	SP
6	1035	-23.5416	-46.6416	sao paulo	SP

8.3 Order-Items Table

1	order_id	order_iten	product_ic	seller_id	shipping_limit_date	price	freight_value
2	00010242f	1	4244733e0	48436dade	19-09-2017 09:45	58.9	13.29
3	00018f77f	1	e5f2d52b8	dd7ddc04e	03-05-2017 11:05	239.9	19.93
4	000229ec3	1	c777355d1	5b51032ec	18-01-2018 14:48	199	17.87
5	00024acbc	1	7634da152	9d7a1d34a	15-08-2018 10:10	12.99	12.79
6	00042b26c	1	ac6c3623c	df560393f	13-02-2017 13:57	199.9	18.14

8.4 Order-Payments Table

1	order_id	payment_sequential	payment_type	payment_insta	payment_value
2	b81ef226f3fe178	1	credit_card	8	99.33
3	a9810da82917af	1	credit_card	1	24.39
4	25e8ea4e93396b	1	credit_card	1	65.71
5	ba78997921bbcd	1	credit_card	8	107.78
6	42fdf880ba16b47	1	credit_card	2	128.45

8.5 Order-Reviews Table

	review_id	order_id	review_score	review_commer	review_comment_	review_creation_da	review_answer_timestamp
2	7bc2406110b926393a	73fc7af87114b39	4			18-01-2018 00:00	18-01-2018 21:46
3	80e641a11e56f04c1a	a548910a1c61477	5			10-03-2018 00:00	11-03-2018 03:05
4	228ce5500dc1d8e020	f9e4b658b201a9f	5			17-02-2018 00:00	18-02-2018 14:36
5	e64fb393e7b32834bb	658677c97b385a9	5		Recebi bem antes	21-04-2017 00:00	21-04-2017 22:02
6	f7c4243c7fe1938f181	8e6bfb81e283fa7	5		ParabÃ©ns lojas la	01-03-2018 00:00	02-03-2018 10:26

8.6 Order Table

	order_id	customer_	order_status	order_purchase_time	order_approved_at	order_delivered_carr	order_delivered_custo	order_estimated_delivery_date
2	e481f51cb	9ef432eb6	delivered	02-10-2017 10:56	02-10-2017 11:07	04-10-2017 19:55	10-10-2017 21:25	18-10-2017 00:00
3	53cdb2fc8	b0830fb47	delivered	24-07-2018 20:41	26-07-2018 03:24	26-07-2018 14:31	07-08-2018 15:27	13-08-2018 00:00
4	47770eb9	41ce2a54c	delivered	08-08-2018 08:38	08-08-2018 08:55	08-08-2018 13:50	17-08-2018 18:06	04-09-2018 00:00
5	949d5b44	f88197465	delivered	18-11-2017 19:28	18-11-2017 19:45	22-11-2017 13:39	02-12-2017 00:28	15-12-2017 00:00
6	ad21c59c0	8ab97904	delivered	13-02-2018 21:18	13-02-2018 22:20	14-02-2018 19:46	16-02-2018 18:17	26-02-2018 00:00

8.7 Products Table

	product_ic	product_c	product_n	product_d	product_p	product_w	product_le	product_h	product_width_cm
2	1e9e8ef04	perfumaria	40	287	1	225	16	10	14
3	3aa07113	artes	44	276	1	1000	30	18	20
4	96bd76ec8	esporte_la	46	250	1	154	18	9	15
5	cef67bcfe	bebes	27	261	1	371	26	4	26
6	9dc1a7de	utilidades	37	402	4	625	20	17	13

8.8 Sellers Table

	seller_id	seller_zip_	seller_city	seller_state
2	3442f8959	13023	campinas	SP
3	d1b65fc7d	13844	mogi guac	SP
4	ce3ad9de9	20031	rio de jane	RJ
5	c0f3eea2e	4195	sao paulo	SP
6	51a04a8a6	12914	braganca p	SP

8.9 Products Category Table

	product_category_name	product_category_name_english
2	beleza_saude	health_beauty
3	informatica_acessorios	computers_accessories
4	automotivo	auto
5	cama_mesa_banho	bed_bath_table
6	moveis_decoracao	furniture_decor

9 DATA DICTIONARY

9.1 Customer Table

a. **customer_id:**
   i. Description: A unique identifier assigned to each customer.
   ii. Data Type: String or Integer (depends on the data format).

b. **customer_unique_id:**
   i. Description: A unique identifier assigned to each customer, usually generated by the e-commerce platform to distinguish between individual customers.
   ii. Data Type: String.

c. **customer_zip_code_prefix:**
   i. Description: The numerical prefix of the customer's zip code, representing the area or region where the customer resides.
   ii. Data Type: String or Integer (depends on the data format).

d. **customer_city:**
   i. Description: The name of the city where the customer resides.
   ii. Data Type: String.

e. **customer_state:**
   i. Description: The name of the state where the customer resides.
   ii. Data Type: String.

9.2 Geolocation Table

a. **geolocation_zip_code_prefix:**
   i. Description: The numerical prefix of the zip code associated with a geographical location.
   ii. Data Type: String or Integer (depends on the data format).

b. **geolocation_lat:**
   i. Description: The latitude coordinate of the geographical location.
   ii. Data Type: Float or Decimal (numeric data type representing latitude).

c. **geolocation_lng:**
   i. Description: The longitude coordinate of the geographical location.
   ii. Data Type: Float or Decimal (numeric data type representing longitude).

d. **geolocation_city:**

i. Description: The name of the city associated with the geographical location.
ii. Data Type: String.

e. geolocation_state:

i. Description: The name of the state associated with the geographical location.
ii. Data Type: String.

9.3 Order-Items Table

a. order_id:

i. Description: A unique identifier for each order placed by customers.
ii. Data Type: String or Integer (depends on the data format).

b. order_item_id:

i. Description: A unique identifier for each item within an order. In cases where an order contains multiple items, each item is assigned a separate order item ID.
ii. Data Type: String or Integer (depends on the data format).

c. product_id:

i. Description: A unique identifier for each product included in an order.
ii. Data Type: String or Integer (depends on the data format).

d. seller_id:

i. Description: A unique identifier for the seller or vendor who fulfilled the order.
ii. Data Type: String or Integer (depends on the data format).

e. shipping_limit_date:

i. Description: The deadline or cutoff date by which the seller is required to ship the order.
ii. Data Type: Date or Date-Time (depends on the data format).

f. price:

i. Description: The price of the product listed in the order.
ii. Data Type: Numeric (Float or Decimal, depending on the currency and precision).

g. freight_value:

i. Description: The shipping cost or freight value associated with delivering the product to the customer.
ii. Data Type: Numeric (Float or Decimal, depending on the currency and precision).

9.4 Order-Payments Table

a. **order_id:**
   i. Description: A unique identifier for each order placed by customers.
   ii. Data Type: String or Integer (depends on the data format).

b. **payment_sequential:**
   i. Description: A sequential number indicating the order of payments within an order. In cases where an order has multiple payments, each payment is assigned a sequential number.
   ii. Data Type: Integer.

c. **payment_type:**
   i. Description: The method of payment used for the order, such as credit card, debit card, bank transfer, etc.
   ii. Data Type: String.

d. **payment_installments:**
   i. Description: The number of installments or payments used to pay for the order.
   ii. Data Type: Integer.

e. **payment_value:**
   i. Description: The total value of the payment made for the order, including any additional fees or charges.
   ii. Data Type: Numeric (Float or Decimal, depending on the currency and precision).

9.5 Order-Reviews Table

a. **review_id:**
   i. Description: A unique identifier for each review submitted by a customer.
   ii. Data Type: String or Integer (depends on the data format).

b. **order_id:**
   i. Description: The identifier of the order associated with the review.
   ii. Data Type: String or Integer (depends on the data format).

c. **review_score:**
   i. Description: The rating given by the customer for their satisfaction with the order experience, typically ranging from 1 to 5.
   ii. Data Type: Integer.

d. **review_comment_title:**

i. Description: The title or summary of the customer's review comment.
ii. Data Type: String.

e. review_comment_message:
i. Description: The detailed comment or feedback provided by the customer along the review score.
ii. Data Type: String.

f. review_creation_date:
i. Description: The date and time when the review was created or submitted by the customer.
ii. Data Type: Date-Time.

g. review_answer_timestamp:
i. Description: The timestamp indicating when the review was answered or responded to by the seller or platform.
ii. Data Type: Date-Time.

9.6 Order Table

a. order_id:
i. Description: A unique identifier for each order placed by customers.
ii. Data Type: String or Integer (depends on the data format).

b. customer_id:
i. Description: A unique identifier for each customer who placed an order.
ii. Data Type: String or Integer (depends on the data format).

c. order_status:
i. Description: The status of the order, indicating its current stage in the fulfillment process (e.g., pending, shipped, delivered).
ii. Data Type: String.

d. order_purchase_timestamp:
i. Description: The timestamp when the order was placed by the customer.
ii. Data Type: Date-Time.

e. order_approved_at:
i. Description: The timestamp when the order was approved by the seller or platform.
ii. Data Type: Date-Time.

f. order_delivered_carrier_date:
i. Description: The timestamp when the order was handed over to the carrier or shipping company for delivery.

ii. Data Type: Date-Time.

g. order_delivered_customer_date:

i. Description: The timestamp when the order was delivered to the customer.
ii. Data Type: Date-Time.

h. order_estimated_delivery_date:

i. Description: The estimated delivery date provided to the customer at the time of order placement.
ii. Data Type: Date-Time.

9.7 Products Table

a. product_id:

i. Description: A unique identifier for each product.
ii. Data Type: String or Integer (depends on the data format).

b. product_category_name:

i. Description: The name or category of the product.
ii. Data Type: String.

c. product_name_length:

i. Description: The length of the product name.
ii. Data Type: Integer.

d. product_description_length:

i. Description: The length of the product description.
ii. Data Type: Integer.

e. product_photos_qty:

i. Description: The number of photos associated with the product.
ii. Data Type: Integer.

f. product_weight_g:

i. Description: The weight of the product in grams.
ii. Data Type: Numeric (Float or Decimal, depending on the precision).

g. product_length_cm:

i. Description: The length of the product in centimeters.
ii. Data Type: Numeric (Float or Decimal, depending on the precision).

h. product_height_cm:

i. Description: The height of the product in centimeters.
ii. Data Type: Numeric (Float or Decimal, depending on the precision)

i. **product_width_cm:**
    i. Description: The width of the product in centimeters.
    ii. Data Type: Numeric (Float or Decimal, depending on the precision).

9.8 Sellers Table

a. **seller_id:**
    i. Description: A unique identifier for each seller or vendor.
    ii. Data Type: String or Integer (depends on the data format).

b. **seller_zip_code_prefix:**
    i. Description: The numerical prefix of the seller's zip code, representing the area or region where the seller is located.
    ii. Data Type: String or Integer (depends on the data format).

c. **seller_city:**
    i. Description: The name of the city where the seller is located.
    ii. Data Type: String.

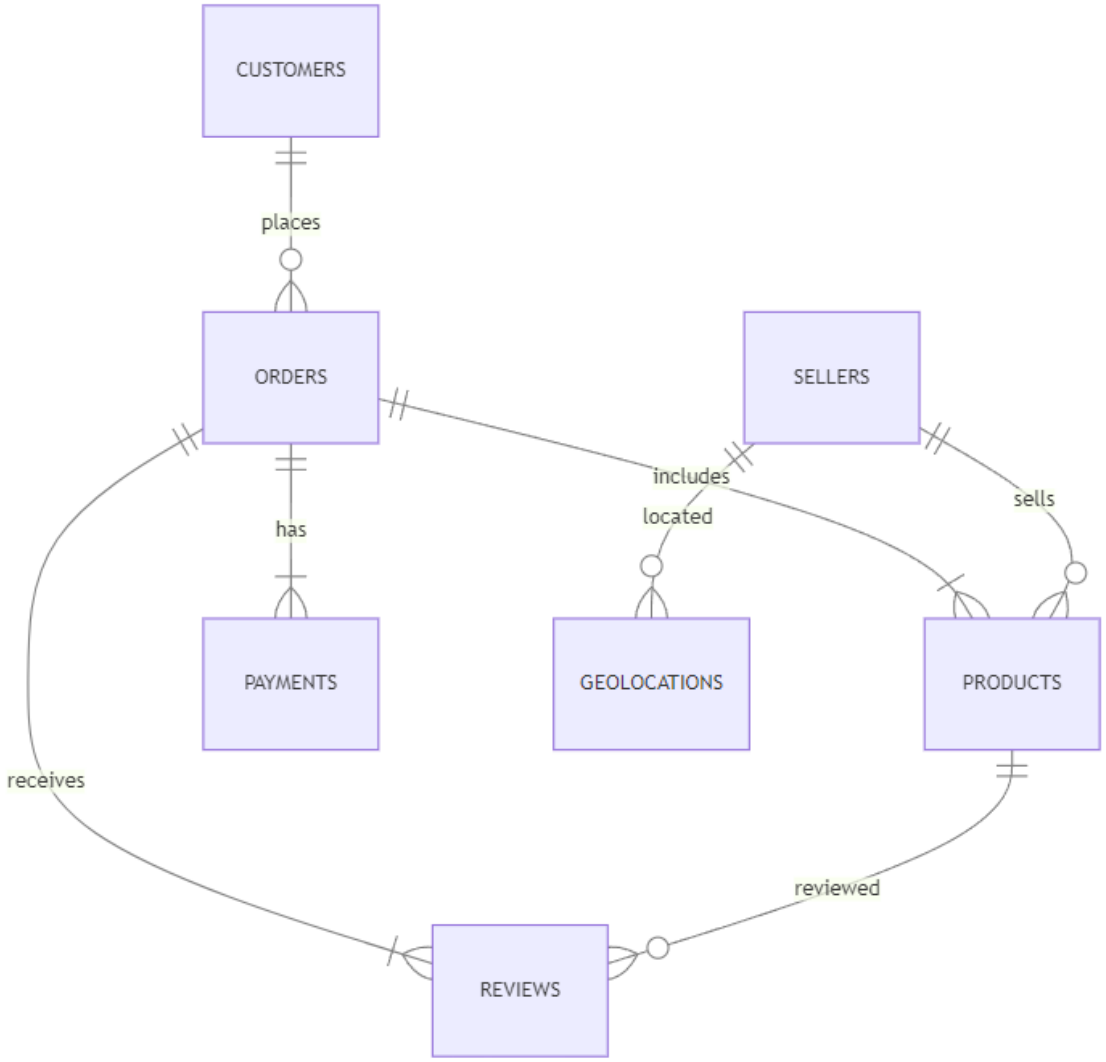
d. **seller_state:**
    i. Description: The name of the state where the seller is located.
    ii. Data Type: String.

9.9 Products Category Table

a. **product_category_name:**
    i. Description: The original name or category of the product in Portuguese.
    ii. Data Type: String.

b. **product_category_name_english:**
    i. Description: The translated name or category of the product in English.
    ii. Data Type: String.

10 RELATIONAL DIAGRAMS



11 PROGRAM DESIGN

Program Design for E-commerce Customer Satisfaction Prediction:

1. Data Preprocessing:

a. Load the dataset
b. Handle missing values
c. Remove duplicates
d. Encode categorical variables
e. Normalize/standardize numerical variables
f. Split the dataset into training and testing sets

2. Exploratory Data Analysis (EDA):

a. Analyze the distribution of positive and negative review scores
b. Investigate the relationship between review scores and various features (e.g., delivery time, price, etc.)
c. Visualize the data using appropriate plots and charts

3. Feature Engineering and Selection:

a. Create new features based on domain knowledge and EDA insights
b. Perform feature selection using statistical methods (e.g., correlation analysis, chi-square test, etc.) or machine learning algorithms (e.g., recursive feature elimination, Lasso regression, etc.)
c. Reduce the dimensionality of the dataset if necessary

4. Model Selection and Evaluation:

a. Choose appropriate machine learning algorithms (e.g., logistic regression, decision trees, random forests, etc.)
b. Train the selected models on the training dataset
c. Evaluate the performance of the models on the testing dataset using metrics such as accuracy, precision, recall, F1-score, ROC-AUC, etc.
d. Select the best-performing model based on the evaluation results

5. Hyperparameter Tuning:

a. Fine-tune the hyperparameters of the selected model using techniques such as grid search, random search, or Bayesian optimization
b. Train the model with the optimized hyperparameters
c. Evaluate the performance of the optimized model

6. Deployment:

a. Integrate the trained model into the e-commerce platform
b. Develop an API or a user interface for the model to receive input data and provide predictions
c. Monitor the performance of the model in production and retrain it periodically with new data

7. Maintenance and Monitoring:

a. Collect feedback from users and stakeholders to improve the model
b. Monitor the performance of the model over time and identify any potential issues or biases
c. Retrain the model with updated data and features as needed
d. Continuously evaluate and improve the model based on new insights and requirements.

12 TESTING

a. Unit Testing:

Test individual components or functions in isolation to ensure they produce the expected output for various inputs.

b. Integration Testing:

Test the interactions between different modules or components to ensure they work together correctly.

c. End-to-End Testing:

Test the entire application flow from start to finish to ensure all components work together as expected.

d. Security Testing

Identify and address potential security vulnerabilities in your application, such as SQL injection, cross-site scripting (XSS), or authentication issues.

e. User Acceptance Testing (UAT):

Involve end users or stakeholders in testing the application to ensure it meets their expectations and requirements.

13 CONCLUSIONS

The project aimed to predict customer satisfaction for purchases made from Brazilian e-commerce intermediated by Olist, using a historical dataset of 100,000 orders from 2016 to 2018. The primary focus was on identifying negative review scores, as they are crucial for business improvement.

The analysis revealed that most positive reviews had longer estimated delivery days but were delivered on or before schedule, while most negative review scores had shorter estimated delivery days but were delivered later than expected. Additionally, negative review scores had a product price only 3.26% higher than positive review scores.

Several machine learning models were evaluated, and the K Neighbors model was chosen based on its recall for negative review scores. By eliminating less relevant predictors and

fine-tuning hyperparameters, the overall performance of the algorithm was improved from an accuracy of 66% to 72%. The correct identification of negative review scores increased from 63% to 65%, which is significant for the business.

However, there is still room for improvement in the model. Future work could focus on feature engineering, using a feature selection algorithm, and testing other models. The current solution provides a scalable process that generates time and resource savings, and the next step is to deploy the solution.

In conclusion, the project successfully developed a machine learning model to predict customer satisfaction based on historical data, enabling Olist to proactively address potential issues and improve their e-commerce services.

13 REFERENCE

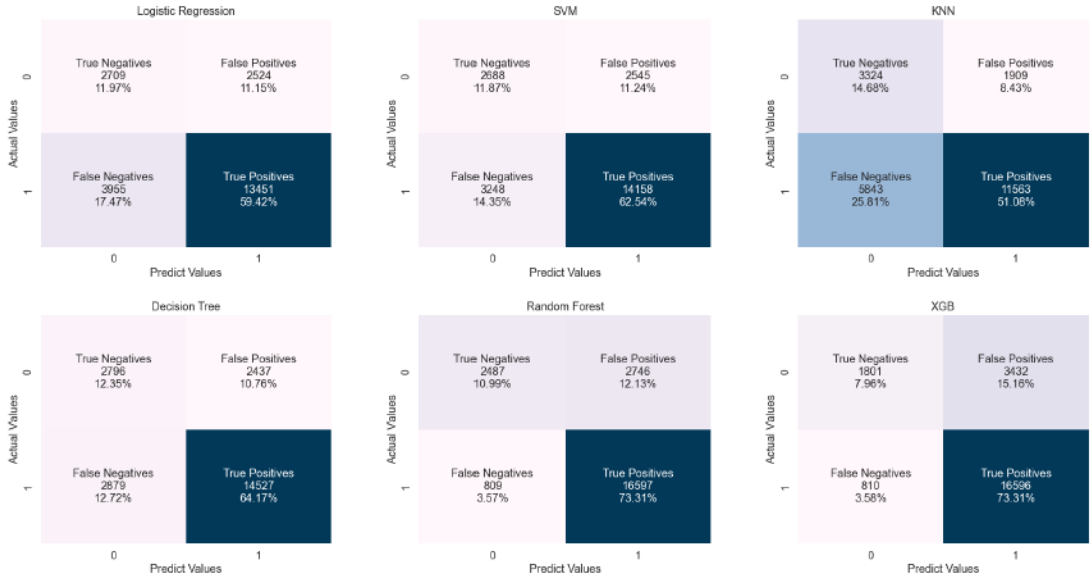
1. Olist. (2018). Brazilian E-Commerce Public Dataset by Olist. Retrieved from https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce
2. Polina G. (n.d.). Illustration. Retrieved from https://icons8.com/illustrations/author/DETSVS1CxEMr
3. Scikit-learn. (n.d.). Machine Learning in Python. Retrieved from https://scikit-learn.org/stable/
4. XGBoost. (n.d.). Scalable, Portable and Distributed Gradient Boosting (GBDT, GBRT or GBM) Library, based on the Gradient Boosting framework. Retrieved from https://xgboost.readthedocs.io/en/latest/
5. Kaggle. (n.d.). Machine Learning Competitions. Retrieved from https://www.kaggle.com/
6. Keras. (n.d.). Deep Learning for humans. Retrieved from https://keras.io/
7. TensorFlow. (n.d.). An end-to-end open source platform for machine learning. Retrieved from https://www.tensorflow.org/
8. Pandas. (n.d.). pandas: a powerful data analysis and manipulation library for Python. Retrieved from https://pandas.pydata.org/
9. NumPy. (n.d.). NumPy: array processing for numbers in Python. Retrieved from https://numpy.org/
10. Matplotlib. (n.d.). Matplotlib: a plotting library for the Python programming language. Retrieved from https://matplotlib.org/
11. Seaborn. (n.d.). Seaborn: statistical data visualization. Retrieved from https://seaborn.pydata.org/

14 SOURCE CODE

PROJECT CODE GITHUB LINK:

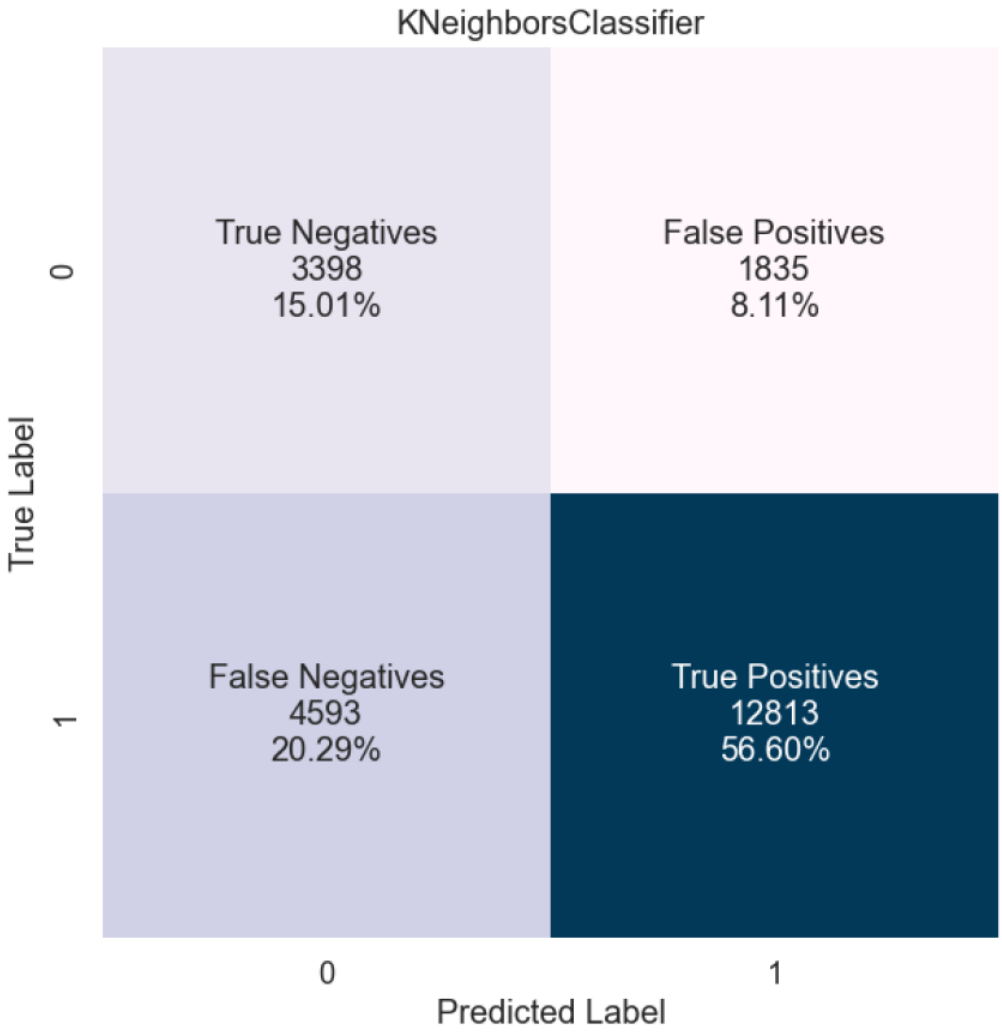
https://github.com/sanjil2/Predicting-Customer-Satisfaction

15 SCREENSHOTS



Out[ ]:

	Model	Accuracy	AUC	Precision (0)	Precision (1)	Recall (0)	Recall (1)	F1-Score (0)	F1-Score (1)	Pre
0	Logistic Regression	0.713812	0.650000	0.406513	0.842003	0.517676	0.772780	0.455409	0.805908	0.6
1	SVM	0.744114	0.660000	0.452830	0.847632	0.513663	0.813398	0.481332	0.830162	0.6
2	KNN	0.657582	0.650000	0.362605	0.858299	0.635200	0.664311	0.461667	0.748947	0.6
3	Decision Tree	0.765184	0.680000	0.492687	0.856343	0.534302	0.834597	0.512651	0.845330	0.6
4	Random Forest	0.842970	0.710000	0.754551	0.858036	0.475253	0.953522	0.583187	0.903263	0.8
5	XGB	0.812624	0.650000	0.689774	0.828640	0.344162	0.953464	0.459204	0.886681	0.7


KNeighborsClassifier
True Negatives
3398
15.01%
False Positives
1835
8.11%
False Negatives
4593
20.29%
True Positives
12813
56.60%
0
1
True Label
0
1
Predicted Label