Summer Internship Project Report

(17th May-8th August 2022)

Gojek

By

Rutvik Kumar

19CE02026

Under the guidance of

Aditya Datta

(Data Science Manager at Gojek)

School of Infrastructure

Indian Institute of Technology Bhubaneswar

TABLE OF CONTENTS

ABSTRACT

In relation to GoFood, on reducing driver subsidy via more efficient logistics, Reliability & Ordering stream will be working on ‘Delay Dispatch’ feature, which delays driver allocation time on each GoFood order dynamically following the food preparation time for each order. This initiative is planned to reduce driver waiting time for each order. Food preparation time modelling has been done before in September 2019. However, on November 2019 GoFood has rolled out Driver Arrival Flow, which can be incorporated in the existing model to increase the performance. We have Develop a model to estimate the time needed by merchant to prepare food for a certain order, considering order time, merchant attributes (location, food type, etc.), and historical food preparation time.

ACKNOWLEDGEMENT

I would like to articulate my profound gratitude and indebtedness to my project guide Aditya Datta who has always been a constant motivation and guiding factor throughout the internship time in and out as well. It has been a great pleasure for me to get an opportunity to work under him and complete the project successfully. I acknowledge my indebtedness to all the crimson iOS team members. Last but not the least, my sincere thanks to all of my friends who have patiently extended all sorts of help for accomplishing this undertaking

PREREQUISITES

1) SQL for Querying and Manipulation:

SQL (Structured Query Language) is a domain-specific language used for managing and manipulating relational databases. Developed in the 1970s, SQL has become a standard for interacting with databases, enabling users to store, retrieve, update, and manage data efficiently. It serves as a bridge between humans and databases, allowing users to perform complex operations using simple syntax.

SQL operates on the principle of a relational database management system (RDBMS), which organizes data into tables with rows and columns. Each table represents an entity, while columns represent attributes of that entity. SQL provides a set of commands to interact with these tables:

1. Data Querying: The `SELECT` statement is used to retrieve data from one or more tables. It allows users to specify which columns to retrieve, conditions for filtering rows, and sorting criteria.
2. Data Manipulation: The `INSERT`, `UPDATE`, and `DELETE` statements modify data within tables. `INSERT` adds new rows, `UPDATE` modifies existing rows, and `DELETE` removes rows.
3. Data Definition: The `CREATE`, `ALTER`, and `DROP` statements manage the structure of tables. `CREATE` generates new tables, `ALTER` modifies existing table structures, and `DROP` deletes tables.
4. Data Constraints: SQL supports constraints like `PRIMARY KEY`, `FOREIGN KEY`, `UNIQUE`, and `CHECK` that ensure data integrity and enforce relationships between tables.
5. Data Joining: SQL allows users to combine data from multiple tables using `JOIN` operations. This helps in querying information spread across different tables.

6. Data Aggregation: The `GROUP BY` clause, along with aggregate functions like `SUM`, `COUNT`, `AVG`, etc., permits the summarization of data based on specified criteria.
7. Data Views: SQL enables the creation of virtual tables known as views. These are based on the results of SELECT queries and provide a simplified way to access complex data.
8. Transactions: SQL supports transactions, allowing multiple SQL statements to be executed together as a single unit. This ensures data consistency and integrity.
9. Subqueries: SQL allows embedding one query inside another, facilitating more complex and dynamic queries.
10. Security and Access Control: SQL supports user authentication and authorization, ensuring that only authorized users can access, modify, or delete data.

Despite its widespread usage, SQL has evolved over time, leading to various dialects like MySQL, PostgreSQL, Oracle SQL, and Microsoft SQL Server. These dialects may have their own specific features and variations. In conclusion, SQL plays a pivotal role in data management and manipulation. Its intuitive syntax and powerful capabilities have made it an indispensable tool for developers, analysts, and administrators working with relational databases. Whether you're a beginner or an experienced professional, a strong understanding of SQL is essential for effectively working with data in a variety of applications.

2) Machine Learning:

Machine Learning (ML) is a subset of artificial intelligence (AI) that focuses on the development of algorithms and models that enable computers to learn from data and make predictions or decisions without being explicitly programmed. It empowers computers to recognize patterns, extract insights, and improve performance over time through experience. Machine learning has witnessed significant advancements in

recent years, transforming industries and shaping various aspects of modern life.

At the heart of machine learning is the concept of learning from data. This process involves three main components:

1. Data: Quality data is the foundation of successful machine learning. Large, diverse, and well-structured datasets are used to train models. These datasets contain input features and corresponding target values or outcomes that the model should predict.
2. Algorithms: Machine learning algorithms are mathematical techniques that learn patterns and relationships from data. They range from simple algorithms like linear regression to complex ones like neural networks. These algorithms adapt and refine their internal parameters based on the data they are exposed to.
3. Model Evaluation and Improvement: After training on data, models need to be evaluated for their performance using metrics like accuracy, precision, recall, and F1 score, among others. Models can then be fine-tuned or optimized by adjusting hyperparameters, adding regularization, or changing the architecture.

Machine learning is broadly categorized into three types:

1. Supervised Learning: In this approach, models are trained on labeled data, meaning the input features are accompanied by the correct output or target. The model learns to map inputs to outputs and can make predictions on new, unseen data.
2. Unsupervised Learning: Here, the model is trained on unlabeled data and aims to find patterns or structures within the data. Clustering and dimensionality reduction are common tasks in unsupervised learning.
3. Reinforcement Learning: This involves training models through trial and error. Models learn by interacting with an environment and receiving feedback in the form of rewards or penalties. They gradually improve their actions to maximize rewards.

The field of machine learning continues to evolve, with advancements in deep learning, which involves complex neural networks, and explainable AI, which aims to make models more transparent and interpretable. As machine learning continues to grow, its integration with various industries promises to drive innovation, improve efficiency, and revolutionize the way we interact with technology.

PROBLEM STATEMENT

Create a model to estimate the time needed by merchant to prepare food for a certain order, considering order time, merchant attributes (location, food type, etc.), and historical food preparation time.

Possible use cases on product implementation:

- GoFood Pick Up ETA (existing, shown on order page)
- GoFood Delay Dispatch (to be built, for back end allocation trigger after order placed)

SOLUTION

1)Path to be Followed

1. Predicting food preparation time using item_list feature
    - Transforming dish item names to vectors. Here are the few methods we are planning to use the item features:
        1. Use PCA to map the 250 item vector to 65 features and then use this item feature.
        2. Take the root square sum for these 250 vectors to create 1 value (magnitude) per item and the corresponding angle.
        3. Use tf-idf to create one value per item
    - Using root dish name to get the subclass value and then map them back to the order number.
      (NOTE: instead of using this feature for every individual item, we will take the average of all and have just one value for each order.)
2. Create bucketing of food preparation time (prediction on time bucket instead of numerical value)
    - Bucketing with 5 buckets between min, 25%, 50%, 75%, 90% and above 90%.

Bucket	**Food Preparation Time**	**Reference**
Bucket 1	0-4 minutes	Percentile 25
Bucket 2	4-8 minutes	Percentile 50
Bucket 3	8-12 minutes	Percentile 75
Bucket 4	12-18 minutes	Percentile 90
Bucket 5	Above 18 minutes	

    - Bucketing with 7 buckets between min, 15%, 30%....., and 90% and above 90%.

- Bucketing with 10 buckets between min, 10%, 20%....., 80%, and 90% and above 90%.

3. Creating model using FPT-realtime and historical features
   - Using historical features only
   - Using both real-time and historical features
4. Using XGBoost and ANN for modelling
   - XGBoost Regression for continuous FPT.
   - XGBoost Classifier for bucketed FPT.
   - ANN(optional)

2)Pre-Analysis

1. HeatMap: A heatmap for null values is a graphical representation that visually highlights the missing or null values within a dataset. It uses colors to indicate the presence or absence of data. In a heatmap, each cell corresponds to a data point, and the color intensity reflects the completeness of the data. Missing values are often represented by a distinct color, making it easy to identify patterns of data gaps. Heatmaps for null values are especially useful for data quality assessment, identifying areas with high missingness, and aiding in data preprocessing decisions before analysis.

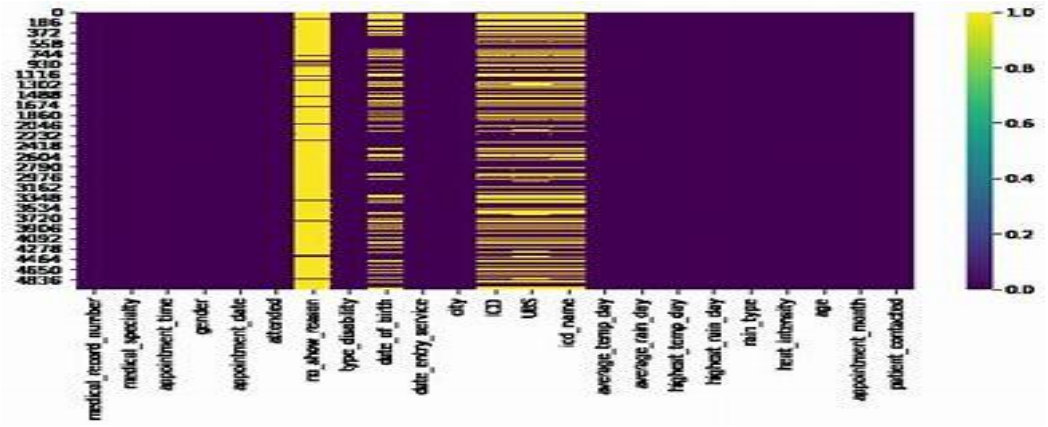

Fig.6.2.1 Heatmap for null values

2. Correlation plot: A correlation heatmap is a visual tool that displays the strength and direction of relationships between

variables in a dataset. Each cell in the heatmap represents the correlation coefficient between two variables, with colors indicating the degree of correlation. Darker colors indicate stronger positive correlations, while lighter colors represent weaker or negative correlations. Correlation heatmaps help identify patterns and connections within data, aiding in feature selection, understanding multicollinearity, and guiding data-driven decisions. They're particularly useful in fields like statistics, finance, and data science for uncovering insights from complex datasets.

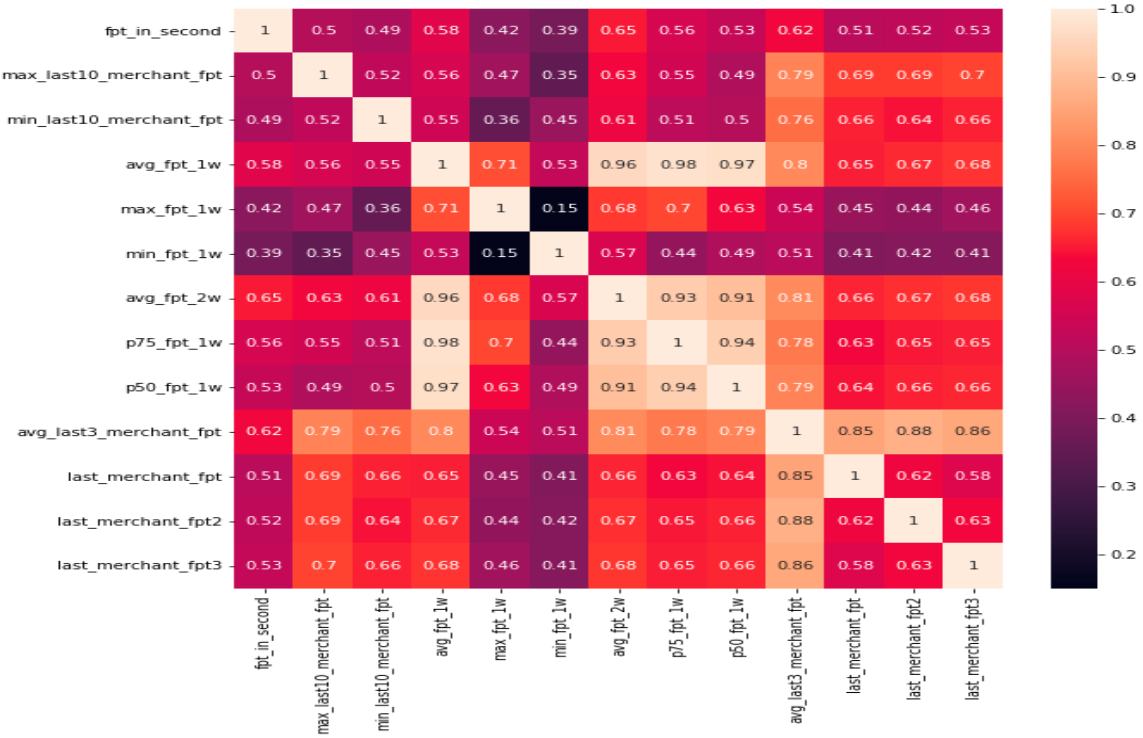

Fig6.2.2 Correlation Plot

3) Modelling and Evaluation

1. Model for Prediction

   There are two major options explored for data modelling, which are using **XGBoost Regressor** and **XGBoost Classifier**.

   1. XGBOOST Regressor: XGBoost (Extreme Gradient Boosting) regression is a powerful machine learning algorithm renowned for its exceptional predictive

performance in regression tasks. It's an ensemble learning method that combines the strengths of decision trees and gradient boosting. XGBoost builds a series of decision trees sequentially, each tree aiming to correct the errors of the previous ones, resulting in a more accurate and robust prediction. XGBoost has proven effective across various domains, including finance, healthcare, and marketing, where accurate predictions are essential. Its robustness, speed, and ability to handle complex relationships in data have contributed to its popularity among data scientists and machine learning practitioners. However, like any algorithm, proper parameter tuning and understanding of its functioning are crucial for optimal results.

2. XGBOOST Classifier: XGBoost (Extreme Gradient Boosting) classifier is a highly effective machine learning algorithm known for its exceptional performance in classification tasks. It's part of the ensemble learning family and is particularly well-regarded for its ability to handle complex relationships within data, making it a popular choice for a wide range of applications. With its robustness, speed, and ability to handle high-dimensional and noisy data, the XGBoost classifier has gained significant popularity among data scientists and machine learning practitioners. Its success stories span domains such as finance, healthcare, and natural language processing. However, achieving optimal results requires tuning hyperparameters and understanding how the algorithm functions in different contexts.

2. Evaluation Metrics
    1. MAE: Mean Absolute Error (MAE) is a common metric used to measure the accuracy of a predictive model. It quantifies the average magnitude of errors between predicted and actual values. MAE calculates the absolute differences between each

prediction and its corresponding true value, then averages these differences. Unlike other metrics that square the errors, such as Mean Squared Error, MAE provides a linear measure, making it more interpretable. Lower MAE values indicate better model performance, with 0 indicating a perfect match between predictions and actual values. MAE is particularly useful in scenarios where all errors should be treated equally and their magnitudes matter.

2. RMSE: Root Mean Squared Error (RMSE) is a common metric used to measure the accuracy of a predictive model. It quantifies the average magnitude of the differences between predicted values and actual values. RMSE calculates the square root of the average of squared differences, providing insight into how well the model's predictions align with the true values. Lower RMSE values indicate better model performance, with zero indicating perfect predictions. It's widely used in various fields, including regression analysis, machine learning, and statistics, to evaluate the quality of predictive models and compare different algorithms in terms of their predictive accuracy.

3. MAPE: The Mean Absolute Percentage Error (MAPE) is a metric used to evaluate the accuracy of a forecasting or prediction model. It measures the average percentage difference between the predicted values and the actual values. MAPE is calculated by taking the absolute percentage difference for each data point, averaging them, and expressing the result as a percentage. MAPE provides insight into the magnitude of errors in the predictions and is commonly used in fields like economics, supply chain management, and finance to assess the quality of forecasts and models.

4. Precision: Precision is an essential evaluation metric for classification models. It measures the proportion of correctly

predicted positive instances out of all instances predicted as positive. In other words, it quantifies the model's accuracy in identifying true positives while minimizing false positives. Precision is particularly valuable when the cost of false positives is high, such as in medical diagnoses or fraud detection. A high precision indicates fewer false positives and better reliability in positive predictions. However, precision doesn't consider false negatives, and its balance with recall is crucial to comprehensively assess a classifier's performance.

5. Recall: Recall, an essential evaluation metric for classification models, measures a model's ability to correctly identify all relevant instances within a dataset. Also known as the True Positive Rate, it focuses on minimizing false negatives, making it especially important in scenarios where missing positive instances has significant consequences. Mathematically, Recall is calculated as the ratio of True Positives to the sum of True Positives and False Negatives. A higher Recall indicates better performance in capturing all positives, but it can be at the expense of increased false positives. Achieving a balance between Recall and other metrics like Precision is crucial, depending on the specific application's requirements.

Learning through this internship

Technical skills

- Swift, memory management (ARC), python, Machine learning and SQL.
- Built sample project, understood about reading and writing to the py files.
- Learned about Model architecture, unit testing, SOLID principles and other design patterns (factory, singleton, dependency injection, delegate etc.).
- Started contributing to the GoFood app and understood the process of deployment (making Jira cards, raising PR, resolving PR comments and merging PR etc.). **Soft skills**
- Made habit of attending every meeting (daily standups, lot of catch ups with Aditya, crimson tech discussions, Townhall, MBR, QBR, crimson virtual coffee etc.).
- Understood the working flow of corporate industry.

References:

- https://www.gojek.com
- https://developer.apple.com/documentation/uikit
- https://www.hackingwithswift.com/quick-start/swiftui
- https://forums.swift.org/
- https://developer.apple.com/videos/wwdc2022/
- https://www.youtube.com/c/CodeWithChris