High Level Design (HLD)

Stores Sales Prediction

-Derrick T

Document Version Control

Date Issued	Version	Description	Author
19-07-2023	1.0	Created	Derrick T

Contents

Abstract:

Nowadays, shopping malls and Big Marts keep track of individual item sales data in order to forecast future client demand and adjust inventory management. In a data warehouse, these data stores hold a significant amount of consumer information and particular item details. By mining the data store from the data warehouse, more anomalies and common patterns can be discovered.

This project focuses on predicting the sales based on multiple factors.

1 Introduction

1.1 Why this High-Level Design Document?

The purpose of this High-Level Design (HLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding, and can be used as a reference manual for how the modules interact at a high level.

The HLD will:

- Present all of the design aspects and define them in detail
- Describe the user interface being implemented
- Describe the hardware and software interfaces
- Describe the performance requirements
- Include design features and the architecture of the project
- List and describe the non-functional attributes like:
    - Security
    - Reliability
    - Maintainability
    - Portability
    - Reusability
    - Application compatibility
    - Resource utilization
    - Serviceability

1.2 Scope

The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system.

2 General Description

2.1 Product Perspective

The stores sales prediction solution is a basically a prediction model (Multiple ML models tried and tested). This solution will help us predict the sales for the upcoming years based on the data (past data).

2.2 Problem statement

Nowadays, shopping malls and Big Marts keep track of individual item sales data in order to forecast future client demand and adjust inventory management. In a data warehouse, these data stores hold a significant amount of consumer information and particular item details. By mining the data store from the data warehouse, more anomalies and common patterns can be discovered.

2.3 PROPOSED SOLUTION

The propose solution is building a machine model that is trained using past data and evaluated, and the final evaluated model is fed in the unseen data to make predictions, in this case the sales amount is predicted.

2.4 FURTHER IMPROVEMENTS

Stores sales prediction can be made more accurate by reducing irrelevant data and adding relevant data which has higher probability of affecting the sales amount.

2.5 Technical Requirements

There were no technical requirements for developing this solution, the default requirements were a Personal Computer with good computing power.

2.6 Data Requirements

The data required was based on the problem statement and the variables or the dependencies provided.

The client provides the dataset with relevant columns as mentioned in the problem statement.

After the solution is developed, user data of the user's choice will be required.

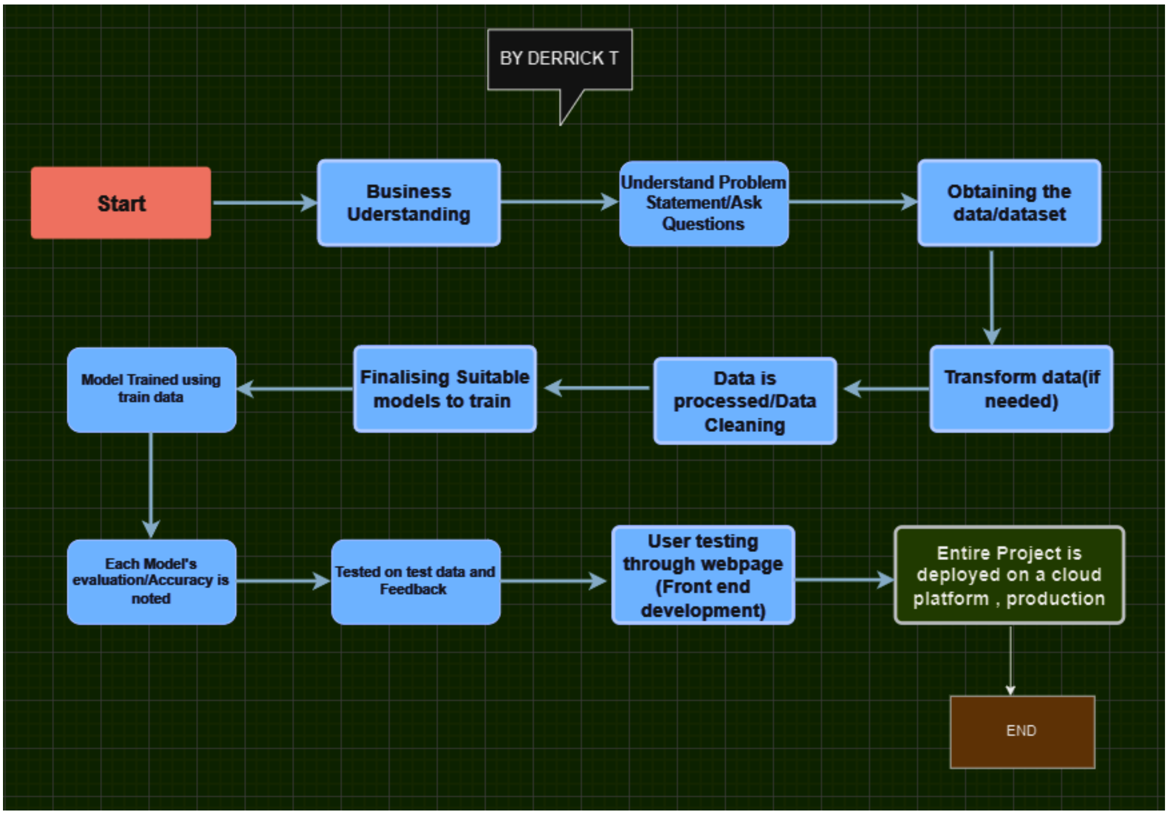
2.7 Tools used

- Python was used as the programming language – to load, pre-process, scale, build ML model and for saving it.

- For visualization, data wrangling etc, frameworks such as NumPy, pandas, matplotlib, scikit learn were used
- PyCharm was used as the IDE for developing a flask application.
- Jupyter notebook as the IDE for developing the solution.
- Visual Studio Code to develop a webpage using HTML and CSS.

3 Design Details

3.1 Process Flow



3.1.1 Model Training and Evaluation

The model was trained using the train data and accuracy was observed on:

For linear regression model:

```
print(mean_absolute_error(y_test,y_pred_lr))
print(np.sqrt(mean_squared_error(y_test,y_pred_lr)))
```

```
888.3515689944207
1174.8379645681518
```

For random forest model:

```
print(r2_score(y_test,y_pred_rf))
print(mean_absolute_error(y_test,y_pred_rf))
print(np.sqrt(mean_squared_error(y_test,y_pred_rf)))
```

```
0.536975573184493
787.945689415224
1123.3478725508144
```

3.1.2 Deployment Process

The entire solution was deployed on a local cloud platform 'Render' linked to GitHub repository which automatically lured the resources needed.

Performance

The solution developed should be very accurate as the sales might differ each time, all the factors when entered correctly by the user will be a step to provide accurate predictions.

Performance can be improved by training the model on new data and evaluating it often.

Reusability

The code written and the components used should have the ability to be reused with no problems.

Application Compatibility

The different components for this project will be using Python as an interface between them. Each component will have its own task to perform, and it is the job of the Python to ensure proper transfer of information.

Resource Utilization

When any task is performed, it will likely use all the processing power available until that function is finished.

Deployment

render

Conclusion

In this machine learning project, we set out to predict store sales using a dataset containing historical sales data, store information, and other relevant features. Throughout the project, we explored and implemented various machine learning techniques to build and evaluate predictive models. The main goal was to create a robust and accurate model that could assist store owners and managers in forecasting sales and making informed business decisions.

Key Findings:

1. Data Preprocessing: We started by carefully cleaning and preprocessing the dataset, handling missing values, and encoding categorical variables. Additionally, we performed feature engineering to extract relevant information from the available data, which helped improve the performance of the models.

2. Exploratory Data Analysis (EDA): Through EDA, we gained valuable insights into the relationships between different features and the target variable (store sales). This analysis enabled us to identify correlations, trends, and patterns, which guided our feature selection process and provided a better understanding of the data.

3. Model Selection: We experimented with various machine learning algorithms, including linear regression, decision trees, random forests. After thorough evaluation and comparison, we identified the best-performing models based on metrics such as mean squared error (MSE), mean absolute error (MAE), and R-squared.

4. Performance Evaluation: To ensure the reliability of our models, we utilized cross-validation techniques and performed hyperparameter tuning. This process allowed us to prevent overfitting and obtain models that generalize well to unseen data.

5. Model Interpretability: Understanding the factors influencing sales predictions is crucial for making informed business decisions.

6. Business Insights: With the deployment of our best-performing model, store owners and managers can now gain valuable insights into their sales forecasting. By using our predictive model, they can anticipate future sales trends, optimize inventory management, plan staffing requirements, and devise effective marketing strategies to drive revenue growth.

Challenges Faced:

1. Limited Data: One of the challenges we encountered was the availability of limited historical sales data. A larger dataset could potentially have improved model performance and generalization.

2. Seasonality and External Factors: Sales data in the retail industry can be influenced by various external factors such as seasonality, economic conditions, and marketing campaigns. Incorporating external data could further enhance the accuracy of our predictions.

3. Feature Engineering: Extracting the most informative features requires domain knowledge and expertise, and in some cases, obtaining additional external data sources may be necessary for more effective feature engineering.

Future Directions:

1. Time Series Models: To better capture the temporal aspects of the data, exploring time series models such as ARIMA, SARIMA, or Prophet could lead to more accurate predictions.

2. Integration of External Data: Incorporating external data sources, such as weather data, social media trends, or economic indicators, may help capture the influence of external factors on store sales.

3. Real-Time Predictions: Implementing the model in a real-time setting would enable store owners to receive up-to-date sales forecasts and respond promptly to changing market conditions.4. A/B Testing: Conducting A/B testing for different marketing strategies and promotions can provide valuable insights into their effectiveness on store sales and further optimize business operations.

In conclusion, our machine learning project on store sales prediction has demonstrated the potential of predictive modelling in the retail industry. By leveraging the power of data and machine learning, store owners and managers can make data-driven decisions, leading to improved efficiency, profitability, and customer satisfaction. Nevertheless, there are opportunities for further improvements and future research to enhance the accuracy and applicability of the models in real-world retail scenarios.