Faculty of Engineering – Cairo University
Credit Hours System – Senior Level
Spring 2024

CMPS451 – Big Data

Project Proposal

Submitted by:

Team 9 – سوبيا الرحمانى

Name	ID	Email
Ahmed Emad Reda	1190180	ahmedemad8@gmail.com
Hla Hany Mohamed	1190344	hla.ahmed00@eng-st.cu.edu.eg
Yomna Osama	1190203	yomna.osamma@gmail.com
Youssef Mohamed Mahmoud	1190202	youssef.shaban01@eng-st.cu.edu.eg

Due Date:

30 March 2024

Idea 1:

The dataset offers a valuable opportunity for businesses operating in the hospitality industry, such as hotels. The objective of this project is to develop a cancellation prediction model using machine learning techniques. For example, if a booking is likely to be cancelled, the hotel can offer the room at a discounted rate to prevent revenue loss. If a high number of cancellations is predicted for a particular date, the hotel can adjust staffing accordingly to avoid overstaffing and reduce costs.

Dataset:

https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand/data

Proposed Solution:

- Preprocess the extracted features by handling missing values, encoding categorical variables, and scaling numerical features to ensure consistency in model training.
- Feature selection techniques can be employed to give weights to relevant features. Unrelated or redundant features can be dropped to improve model performance.
- Perform EDA to gain insights into booking patterns and cancellation rates. Visualize correlations using techniques like histograms, box plots, and heat maps to identify trends within the dataset. Analyze factors such as lead time, arrival date details, guest demographics, and deposit type to understand their impact on cancellation rates.
- Select suitable classification algorithms such as logistic regression, decision trees, random forest or XGBoost for predicting booking cancellations.
- MapReduce can be used to distribute the data across nodes, with mappers calculating necessary statistics, and reducers aggregating values to compute classification probabilities.

Idea 2:

The dataset presents an opportunity for businesses operating in the mobile app industry. The objective of this project is to analyze the Google PlayStore Android app data and derive insights to enhance app performance by predicting app ratings. This prediction can help developers understand how well their app might be received by users even before its release. Market analysis can be conducted on the dataset such as which categories have the highest number of apps. This information can guide businesses in identifying niche markets or popular categories where there may be opportunities for growth or competition, or areas where there may be gaps in the market for new app developments.

Dataset:

The dataset encompasses over 2.3 million app records with 24 columns.

https://www.kaggle.com/datasets/gauthamp10/google-playstore-apps/data

Proposed Solution:

- Preprocess the dataset by handling missing values, encoding categorical variables, and scaling numerical features to ensure consistency in model training. Employ feature selection techniques to prioritize relevant features for app performance prediction.
- Conduct EDA to gain insights into app characteristics and user preferences. Visualize correlations using techniques like histograms, box plots, and heat maps to identify patterns within the dataset. Analyze factors such as app size, pricing, and release date to understand their impact on app ratings and user engagement.
- Utilize regression algorithms such as linear regression, decision trees, random forest, ANN, or XGBoost for rating prediction.
- In MapReduce, the mapper extracts relevant information from the dataset, such as app categories, installs and ratings. The reducer can calculate the average rating and total installs for each app category by aggregating the results from different mapper nodes.