Summary

Demand Prediction for the Santander Bikes across London

This study focuses on predicting the demand and revenue for Santander Cycles in London through a comprehensive analysis of historical usage data, integrating datasets that include cycling patterns and weather conditions. The primary goal is to develop a model that accurately forecasts the demand for Santander bikes and their consequent revenue, a task that is crucial for making informed decisions about bike placements and availability, thereby optimizing operational efficiency and maximizing revenue. Additionally, reliable demand prediction supports efforts to promote a carbon-neutral economy by encouraging the use of eco-friendly transportation. This analysis also examines the proximity of bus and tube stations as a peripheral factor affecting cycle usage through geospatial matching with bus stop and tube station data using BigQuery GIS, aiming to understand the dynamics of urban mobility and how it can be enhanced by integrating different modes of transport.

The methodology involved an extensive exploratory data analysis (EDA) to reveal patterns and correlations. EDA indicated a significant interdependence between cycling and public transportation in London, with a chi-square test highlighting habitual travel routes between start and end stations. Geospatial analysis further showed that cycle stands in central London have high utilization, correlating with their proximity to public transport facilities. Key commuter hotspots were identified around major bus stops and tube stations, suggesting frequent integration of cycling with buses and tubes. These findings underscore the importance of seamless integration between cycling and public transportation in urban planning and transport policies to enhance urban mobility and accessibility.

For modelling purposes, the study used aggregated data from cycle hires along with weather data. The analysis employed several predictive modelling techniques using Python within the PySpark framework to handle large datasets efficiently. Three models were constructed for demand prediction: Linear Regression, Random Forest, and Gradient Boosted Trees (GBT). The GBT model stood out for its precision in forecasting demand due to its proficiency in managing the complex, non-linear relationships inherent in the data. For revenue prediction, time-series models ARIMA (Autoregressive Integrated Moving Average), SARIMA (Seasonal Autoregressive Integrated Moving Average), and LSTM (Long Short-Term Memory network) were used. Among these, the SARIMA model achieved the lowest Root Mean Square Error (RMSE) of 321.350 on the testing data, followed closely by the LSTM model with an RMSE of 327.863, and the ARIMA model with an RMSE of 403.691.

The study concluded that the GBT model was the most effective for demand prediction, leveraging comprehensive data analysis that included weather conditions, public holidays, and proximity to public transportation, thereby capturing seasonal trends and weather fluctuations. For revenue forecasting, the SARIMA and LSTM models, which showed higher accuracy in testing data, predicted a decline in total revenue, while the ARIMA model forecasted stable revenue. However, these forecasts are inherently uncertain and require continuous validation against real-world data. This necessitates ongoing monitoring and refinement of the forecasting models to adapt effectively to changing demand dynamics and optimize resource allocation for the bike rental service.

In summary, this research presents a framework for predicting the demand and revenue for Santander Cycles, offering insights for enhancing urban mobility services. The integration of environmental and temporal variables into the predictive models provides a nuanced understanding of urban transport dynamics, supporting efforts to promote sustainable mobility in London. The findings have significant implications for urban planning and transport policy, suggesting that enhancing the integration of cycling with other public transport modes can improve urban accessibility and reduce energy waste. This study underscores the value of predictive analytics in managing urban transportation systems and promoting eco-friendly commuting options, highlighting the need for continuous model refinement to maintain accuracy and adapt to evolving demand patterns.

Link of the Dataset: https://console.cloud.google.com/bigquery?p=bigquery-public-data&d=london_bicycles&page=dataset&project=angular-stacker-408217&ws=!1m5!1m4!4m3!1sbigquery-public-data!2slondon_bicycles!3scycle_hire Variables: 12 variables in “cycle_hire” dataset and 12 variables in “cycle_stations” dataset. Data Size: ~9.4GB