Feature Engineering with Python Project

Project Deliverables

You will be required to provide the following deliverables.

- A python notebook with your solution.

Instructions

Background

Logistics in Sub-Saharan Africa increases the cost of manufactured goods by up to 320%; while in Europe, it only accounts for up to 90% of the manufacturing cost. Sendy is a business-to-business platform established in 2014, to enable businesses of all types and sizes to transport goods more efficiently across East Africa. The company is headquartered in Kenya with a team of more than 100 staff, focused on building practical solutions for Africa's dynamic transportation needs, from developing apps and web solutions to providing dedicated support for goods on the move.

Problem Statement

Sendy has hired you to help predict the estimated time of delivery of orders, from the point of driver pickup to the point of arrival at the final destination. Build a model that predicts an accurate delivery time, from picking up a package arriving at the final destination. An accurate arrival time prediction will help all business to improve their logistics and communicate the accurate time their time to their customers. You will be required to perform various feature engineering techniques while preparing your data for further analysis.

You will be required to go through the following:

- Defining the Research Question
- Data Importation
- Data Exploration
- Data Cleaning
- Data Analysis (Univariate and Bivariate)
- Data Preparation

- Data Modeling
- Model Evaluation
- Challenging your Solution
- Recommendations / Conclusion

Dataset Information

The dataset provided by Sendy includes order details and rider metrics based on orders made on the Sendy platform. The challenge is to predict the estimated time of arrival for orders- from pick-up to drop-off. The dataset provided here is a subset of over 20,000 orders and only includes direct orders (i.e. Sendy “express” orders) with bikes in Nairobi. All data in this subset have been fully anonymized while preserving the distribution.

Dataset URL = https://bit.ly/3deaKEM

Dataset Glossary = https://bit.ly/30O3xsr

Hint

Create your base model (ensemble regressor), then later improve the accuracy of the base model by performing the following feature engineering techniques:

- Feature improvement
    - Handle categorical features
    - Find and deal with missing values
    - Handle any outliers in your dataset
- Feature scaling (normalisation or standardisation etc.)
- Feature construction
    - Create new features i.e. speed = distance/time, manhattan distance from pick up latitude and Longitude i.e. manhattan distance, haversine distance, bearing, centre point etc.
- Feature Selection
    - Filter methods
    - Feature transformation (PCA, LDA, etc)
    - Wrapper methods

Acknowledgements

Project Source: https://bit.ly/2Y6Hzz3