Machine Learning Project 2

Predicting the Estimated Time of Arrival for Bikes in Nairobi

Despite the continent offering tremendous opportunity, one of the biggest problem is logistics. The complexity of this problem is because of the unique market and the volatile economic drivers. Sendy, in partnership with insight2impact, is hosting a challenge on Zindi to predict the estimated time of delivery of orders, from the point of driver pickup to the point of arrival at final destination. For this project, a training dataset from a subset of about 20,000 orders with bikes (“boda-boda”) in Nairobi is provided. This data is also anonymized. There are also additional GIS dataset that can be leveraged for this project. The description and link to other dataset can be seen here:

https://zindi.africa/competitions/sendy-logistics-challenge/data

For this project, we’re presented the training set (with labels ETA), the test set, riders description and definitions of the variables. The data is anonymized to preserve sensitive information.

Question 1

- Set up a Zindi account (www.Zindi.Africa) and share with the TA’s and instructors. You’ll submit your solutions to Zindi using the sample submission.

Question 2

- Prepare a comprehensive exploratory data analysis (EDA) using the training dataset. The EDA should include joining the different tables (train and riders), exploring the shape of the data, checking correlation matrix, simple and complex plots like heat map and KDE, summary statistics etc.

Question 3

- Check for missing values in the data. How many data are missing? What percentage of the data is missing. Are the missing values categorical or continuous values? Do you think you should input the missing values or just delete the feature or observation for that missing value? How did you get to this conclusion?

Question 4

- What new features did you engineer with the data? Explain the different steps used to engineer and encode the following features: datetime, coordinates, categorical and numerical variables?

Question 5

- Take the training data from the Zindi challenge and split it into 80% train set and 20% test set. Use random seed = 42.
- Train a multiple linear regression on the training data without the features you engineered. What is the training RMSE? What is the MSE when you predict the ETA with the 20% test data?
- Train a multiple linear regression on the training data and include the features you engineered. What is the training RMSE? What is the MSE when you predict the ETA with the 20% test data?

Question 6

- Train a Lasso, Ridge and Elastic Regression on the 80% train set (with the engineered features). What hyperparameters did you use? What is the training RMSE? What is the MSE when you predict the ETA with the 20% test data?

Question 7

- Now choose your best algorithm from multiple linear regression, lasso regression, ridge regression and elastic net. Training this algorithm with 100% training data. Predict the ETA using the test data provided. Upload your solution for the Zindi challenge.

Question 8

- Do you think there are more stuff you can do with your data. Explore and enjoy. This is a bonus question.