**EDA on Airbnb Dataset**

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| Name: Debanjan Ganguly  Email: [officialdeba10@gmail.com](mailto:officialdeba10@gmail.com)  Project Contribution:   * Data Wrangling – Checked for null and duplicated values * Plot Distribution of numeric features * Plot spatial density graph to show frequent routes * Univariate and Bivariate analysis against target * Plot correlation matrix * Feature engineering, encoding of categorical features and standard scaling after splitting * Train and test model with Linear, Lasso and Ridge Regression, DecisionTreeRegressor, Random Forest and LightBGM * Plotting and Evaluation of Metrics of Model |
| **GitHub Repo & Drive Link:** |
| Github Link: - <https://github.com/awesomedeba10/NYC-Taxi-Trip-Duration-Prediction---AlmaBetter-Capstone-Project>  Drive Link: - <https://drive.google.com/drive/folders/16ns27oxBK18ipWj6oah8tV6_emuMAeCc?usp=sharing> |
| **Project Summary** |
| Machine learning has been of significant help as it has helped businesses in abundant ways. We will create a model which will predict taxi trip time duration.  Most on-demand taxi platforms require a way to know the estimated time which driver will be occupied. It is important to predict how long a driver will have his taxi occupied. If a dispatcher or system got estimates about the taxi driver's current ride time, they could better recognize which driver to allocate for each pickup request which results to be less waiting time which means less cancellation from client side and increase in profit margin, as well as customer base.  First thing we noticed the dataset has more than 14 lakhs of data, so we’ve implemented function which will convert dtypes of pandas dataset to numpy.int format to reduce memories. We’ve also noticed the target variable has positive extreme right skewed distribution; we’ve applied log transformation to target variable. We’ve plotted graphs, done both univariate and bivariate analysis, tried to extract general overview and stories of dataset. Also, we’ve created some new fields like distance, month and day name etc. We’ve split the dataset into 3 parts, train, test and validation. Then, we’ve applied StandardScaler by sklearn to normalize the data. We’ve tried Linear, Lasso and Ridge Regression, DecisionTreeRegressor, RandomForestRegressor and LGBMRegressor. Out of these, we found best r2\_score to be ~0.75 on validation dataset using lgbm model. Though the evaluation metrics on prediction is not that much accurate, but we can use some external features like rating by user, driver experience level, driver rating or traffic conditions etc., these can significantly boost our model’s performances. |
|  |