Wolt Assignment

Manish Saraswat
17.04.2023

Workflow

Following is the workflow used in solving the given problem:

1. Start with data exploration: understand the patterns using univariate and multivariate analysis
2. Based on understanding derived from data exploration, now we create/transform features in a format using which we can train some models
3. Based on target variable:
    - decide a validation strategy
    - decide which algorithms can be used for model training
    - decide which metrics to use for model evaluation
4. Analyse model predictions

LIBRARIES/SYSTEM USED

- Operating System: Macos
- Language: Python
- Environment: Conda
- IDE: Jupyter Notebook
- Libraries used:
  - Data Exploration: pandas, seaborn, matplotlib
  - Model training: scikit-learn, catboost
  - Feature Importance: shap

SUMMARY

- Using data exploration, we find:
  - The target variable is binary and imbalanced.
  - The given feature set is anonymised.
- Using feature transformation/engineering, we do:
  - We extract hour, min, seconds from time
  - Create a dummy feature for missing values
  - One hot encode the categorical feature
  - Create an age feature based on time difference
- Model Training
  - Built logistic regression, random forest and catboost models
  - As per F1 score, random forest model outperforms