Data Science
Summer Intern assignment 2024

Insights on daily average of courier partners online

Data Exploration

- After plotting and checking on correlation of our features we observe;
  - The target value (courier partners online) has a seasonality and an upward trend.
  - The temperature and humidity are almost perfectly negatively correlated and follow a seasonal pattern as well.
  - Precipitation is a bit randomly spread with some extreme peaks and few ‘bell curve’ shaped periods.
  - All of the above are important for working on data preprocessing and on choosing an appropriate predictive model.

Data preprocessing

- There seem to be also some minor issues with our data, so I opted for imputing the data accordingly.
  - Courier partners online; Five values seemed out of the medium range (like 10 or 15 times greater) so I guess they are errors in capturing the data rather than normal outliers. I imputed with the median, as it would be less affected by these large outliers
  - Temperature; there are some missing values so I imputed them using linear interpolation
  - Precipitation, again some missing values, filled with the median of this feature.
  - As for last step, was to normalise the data so they would be on the same range (0,1),assisting the models to capture equally the features' influence on target value.

Data split

- The data was split to train-validation-test ratio of 80-10-10.
- The chronological order of the data was maintained.
- Purpose of Splitting:
  - Training Set: Used for training the models.
  - Validation Set: Used for model tuning and hyperparameter optimization.
  - Test Set: Used for final model evaluation to simulate real-world predictions.

XGBoost model

- XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.
- In our case the hyperparameters that were tuned were;
  - max_depth: Controls the maximum depth of the trees.
  - learning_rate: Step size shrinkage used to prevent overfitting.
  - n_estimators: Number of trees in the ensemble.
  - min_child_weight: Minimum sum of instance weight needed in a child node.

Support Vector Regression (SVR)

- Support Vector Regression (SVR) is a type of Support Vector Machine (SVM) used for regression tasks. It's a powerful algorithm in machine learning derived from the concept of support vector machines, a set of supervised learning methods used primarily for classification and regression analyses.
- In our case the hyperparameters that were tuned were;
  - C; controls the trade-off between smooth decision boundary and classifying training points correctly.
  - Epsilon; specifies the epsilon-tube within which no penalty is associated in the training loss function with points predicted within a distance epsilon from the actual value.
  - Kernel; defines the type of kernel function used to transform the input data into a higher dimensional space.

Evaluation and further insights

- Both XGBoost and SVR models have been successful in capturing the patterns in the time series data of courier partners. This was due to the ability of these models to understand complex relationships and seasonality in the data.
- In order to make future predictions we would need to pass on the according features to these models or use another type of deep learning model or a time series one(ARIMA,SARIMA etc) to make predictions based solely on the trend of our target value.
- It is also important to periodically retraining the models with the latest data to maintain prediction accuracy.
- We could also add extra features that may affect our target value and help us make more precise predictions f.e. holidays,