Methodology

- The dataset which I considered has 14,204 records in total.
- From which, I have considered 8523 records for training set and 5681 records for testing set.
- I have split the training set into train and valid sets during training the models. However, did not consider cross validation for the models.
- For the tuning of hyperparameters in Ridge regression, Lasso regression and Decision tree models are taken, and the values considered are:

  For Ridge regression:

  ```
  alphas = [0.01, 0.1, 1, 10, 100]

  cv=5
  ```

  Lasso regression:

  ```
  alphas = [0.1, 1, 10]

  cv=5
  ```

  Decision tree:

  ```
  param_grid = {

    'max_depth': [3, 5, 10, 20, 30],

    'min_samples_split': [2, 5, 10],

    'min_samples_leaf': [1, 2, 4],

    'max_features': ['sqrt', 'log2']

  }
  ```

- These are the parameters considered while tuning the models and to get the best hyperparameter, GridSearchCV technique is used.
- I have performed Linear Regression for the Baseline and further Lasso regression; Ridge regression and Decision tree models are developed.
- Hyperparameters that achieved best results for each model are:

  Ridge : Best Alpha: 0.1
  Lasso: Best alpha: 0.1
  Decision tree:

Best Parameters: {'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2}

Results

Among the developed models, decision tree model has outperformed others and has given the best results.

Results from the considered papers:

Model	Score
Linear regression	1169
Lasso regression	1178
SVM	1046

Results obtained from the developed models:

Model	RMSE score	R2 square score
Linear Regression (Baseline model)	1067.7242	0.5805
Ridge regression	1067.6091	0.5800
Lasso regression	1068.8606	0.5796
Decision Tree model	1502.5768	0.1693

- Considering the above results, we can say that the, the developed models have performed better compared to the results of considered papers.
- Linear regression, ridge regression, and lasso regression models have similar RMSE and R2 square scores, with the decision tree model having a higher RMSE and lower R2 square score.
- However, after hyperparameter tuning, the decision tree model shows an improvement in RMSE and R2 square score.

- Hence, proceeded with the tuned decision tree model on test set.

Discussion

- Feature engineering techniques like one-hot encoding and label encoding helped converting non-numerical features to numerical and Feature scaling has helped to prepare the final dataset for training.
- Initially, decision tree model was having a higher RMSE and lower R2 square score.
- And after hyperparameter tuning, the decision tree model shows an improvement in RMSE and R2 square scores.

Conclusion

- Most of the techniques or the models that are used in the project are based upon classroom learning and help from the assignments. However, there are minor methods using python like replacing null values, filling missing values or usage of lambda functions which I haven’t learned from the classroom teaching.
- The future improvements to the model can be done considering more relevant features for model training or using ensemble methods to improve overall performance.