Milestone 1 : Apartment Rent report

Team : CS-10

Team Members :

كريم فتحي محمد امين

ID : 2021170406

فيلوباتير متي جاد الكريم العبد

ID: 2021170398

عمرو سعيد كامل

ID: 2021170375

يوسف عصام عزت اخنوخ

ID: 2021170638

فادي عماد شنودة سدره

ID: 2021170380

Preprocessing techniques:

1- We dropped the following columns: id , category, title , body, currency, fee , price as category , fee and currency they all have one value so it wont affect the model. Also title and body has information found in other columns so they were redundant, id was dropped as it only contain unique values.

2- We also optimized price_type as it has values like "Monthly|Weekly" which is neither weekly nor monthly so it was removed, And if the value is weekly we changed the value of price_display to (price_display * 4) so it will be rent per month.

3- Price_display was optimized as it has values like ($, ',') which were removed. Also we filtered strings like 'Weekly' from price_display values.

4-We filled null values for categorical columns by the mode and for numerical columns by the mean except longitude and latitude.

5- we applied label encoder for the following columns cityname , state, has_photo , pets_allowed, source.

6- Time column had timestamps values that we changed to actual years.

7- we applied one hot encoder from scratch to the amenities column and dropped amenities.

8- Outliers were removed from price_display.

9- we applied mini-max scaler to the following columns square_feet , state.

Data Analysis:

1-we used correlation matrix to find the correlation between columns and the price_display column.

2- it was found that the following columns bedroom, bathroom, square_feet , state, longitude, latitude, playground, dishwasher , garbage_disposal and cable or satelite had the best relation. 3- so we applied feature selection and choosed those columns and used them for the whole models.

Regression techniques:

-Random Forest model

- data was split 70 % training and 30 % testing
- - we applied Random forest model which has accuracy 76.1% and RMSE 333 and it is the highest model we had.

-XGBOOST model

- data was split 80 % training and 20 % testing
- we applied xg model which has 5 parameters colsample_bytree parameter specifies the fraction of features to be randomly sampled and is setted to 30% , learning rate set to 1 ,

max depth set to 2 , alpha which penalizes the model for large coefficient values set to 5 and estimator set to 900.

- mean square error for the xgboost model was 363.5 and accuracy was 73.7%.

-**Linear regression model:**

1- data was split 70% training and 30% testing for this model.

2-mean square error for linear model was 567.7 and accuracy is 30.8%

-Polynomial regression model:

1- data was split 70% training and 30% testing for this model

2- mean square error of polynomial is 200731

3- this is the worst model used so it was commented in the code.

Differences:

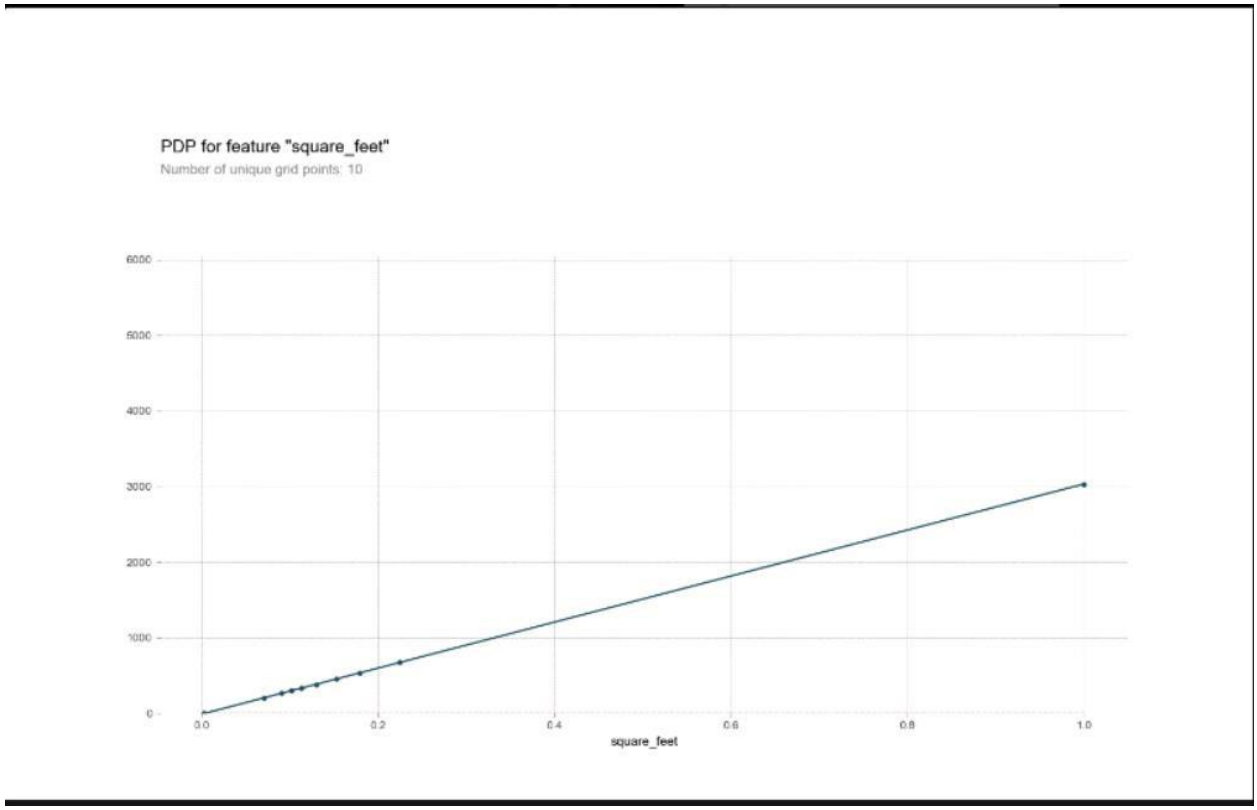
- XGBOOST model is tree based while linear model is not.

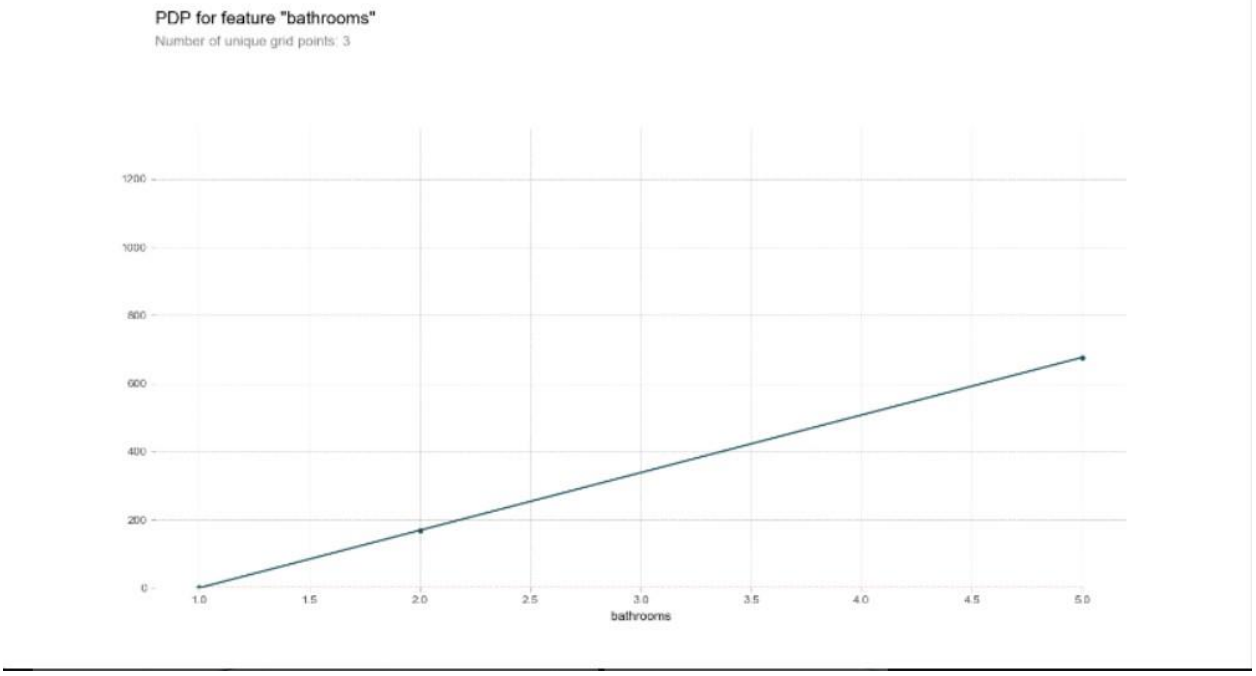
-Linear assume linear relations between feature and target while xgboost does not.

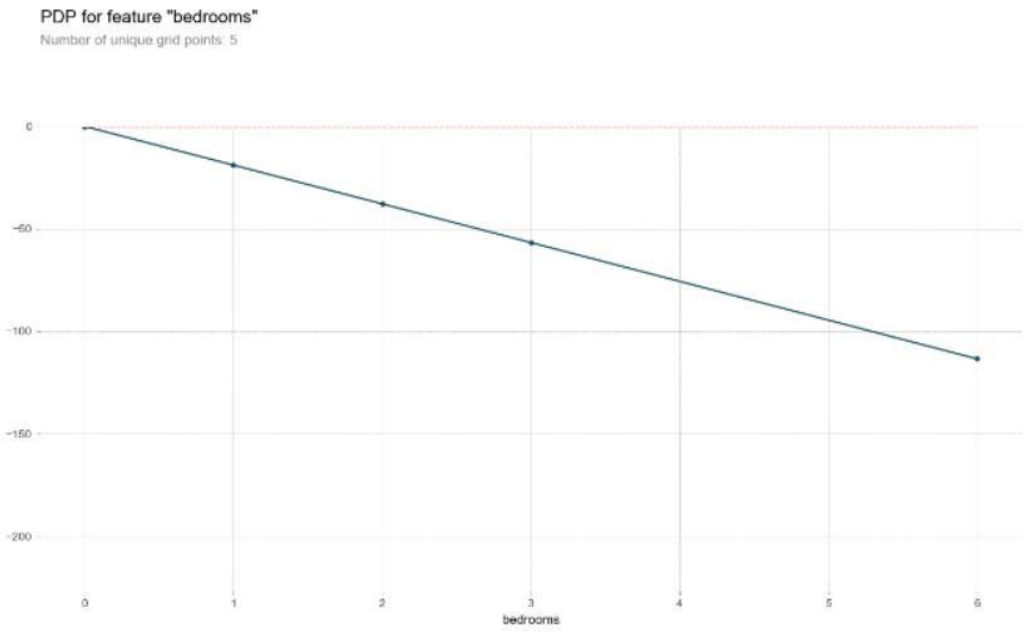
-xgboost had mean square error 363.5 and accuracy was 73.7% while the linear model had mean square error was 567.7 and accuracy was 30.8%.

Regression plots :

Plot for the highly correlated features with the target .


PDP for feature "square_feet"
Number of unique grid points: 10
6000
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0
square_feet


PDP for feature "bathrooms"
Number of unique grid points: 3
1200
1000
800
600
400
200
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
bathrooms

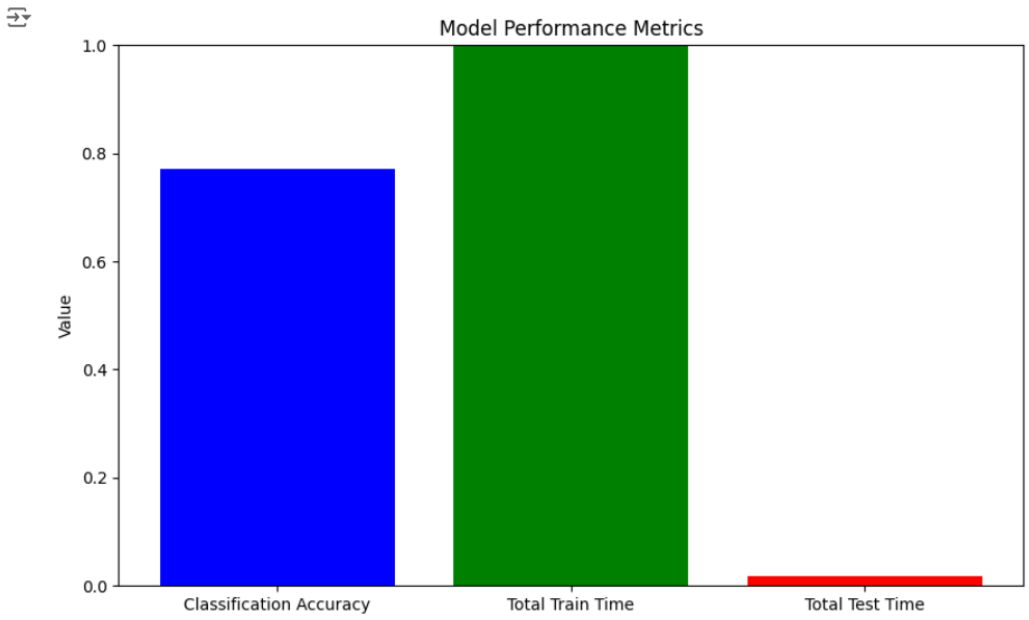


Conclusion:

This problem should be solved using tree based model as it is accurate than other regression models. Our intution was that certain columns like amentites ,bedroom, bathroom will be significant for our model and it was proved after viewing the correlation matrix.

Milestone 2: Apartment Rent report

Classification



The previous figure shows that the Total Train time takes 69.348 sec while the Total test time takes 1.035 sec and Best Classification Training script accuracy is 0.773 by Random forest model .

The feature selection process in Classification is Anova (SelectKBest) And in Regression model is Correlation .

Anova : is used to identify the features most relevant to the target variable in a classification problem by calculating the F-statistic, which represents the ratio of variance between the groups to the variance within the groups. It then selects the top k features with the highest F-scores. It can handle both continuous and categorical features but is mostly used in classification models.

Correlation : is used to identify the linear relationship between independent variables and the dependent variable (target) to indicate how changes in one

variable are associated with changes in another variable. It calculates the correlation coefficient between independent variables and the dependent variable. The coefficient, which ranges from -1 to 1, indicates the strength and direction of the linear relationship

It is disproved because the two feature selection processes yield almost the same best features, indicating that they are capturing similar patterns or relationships within the data.

Hyperparameter Tuning : We used four different models: Random Forest, AdaBoost, Gradient Boosting, and SVM (RBF). In the Random Forest model, we adjusted two hyperparameters: n_estimators (the number of trees in the forest) and max_depth. We performed a GridSearch with 5-fold cross-validation and tried three different values for each hyperparameter: 100, 150, and 200 for n_estimators, and 10, 20, and 30 for max_depth. We found that using 200 trees with a max_depth of 20 yielded the best accuracy of 77.3%.

For the Gradient Boosting model, we manually changed the n_estimators hyperparameter to 100, 150, and 200, and the learning rate to 0.01, 0.1, and 1. We found that using 200 trees with a learning rate of 0.1 resulted in the highest accuracy of 76.5%.

Conclusion : We suspected that using a tree based model gives highest accuracy and it was proved as random forest gives highest accuracy like regression model with accuracy 77.3% .