

## 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 satellite 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.

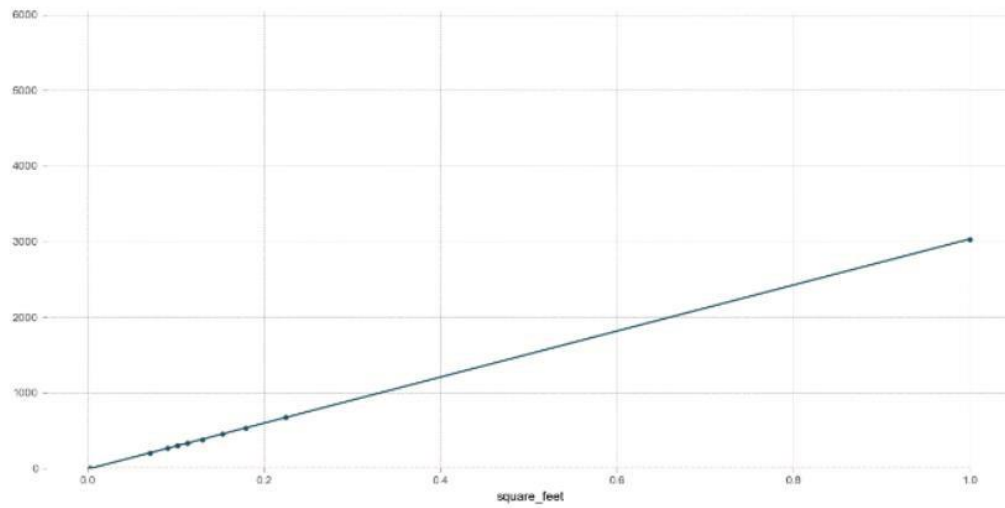
-xgboost had mean square error 363.5 and accuracy was 73.7% while the linear model had mean square error was 567.7 and accuracy was 30.8%.

#### **Regression plots :**

Plot for the highly correlated features with the target .

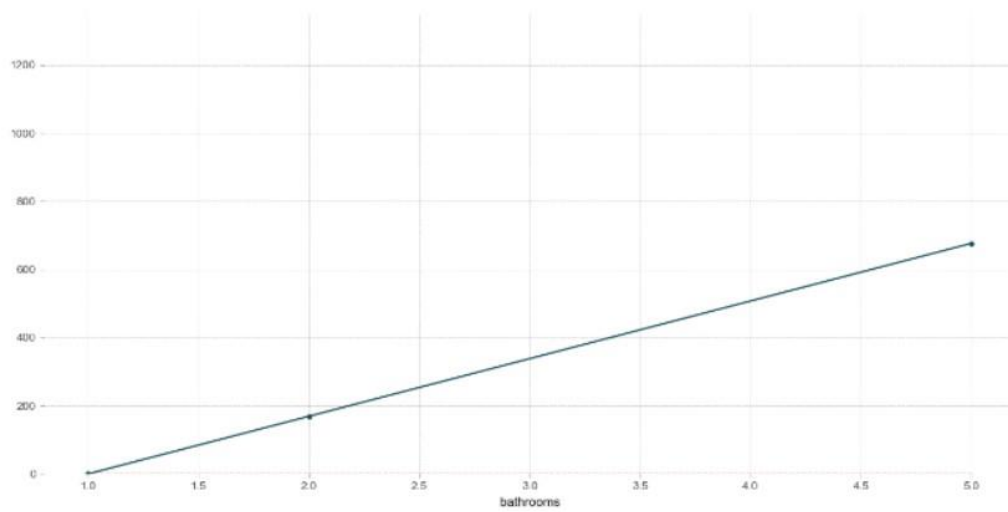
### PDP for feature "square\_feet"

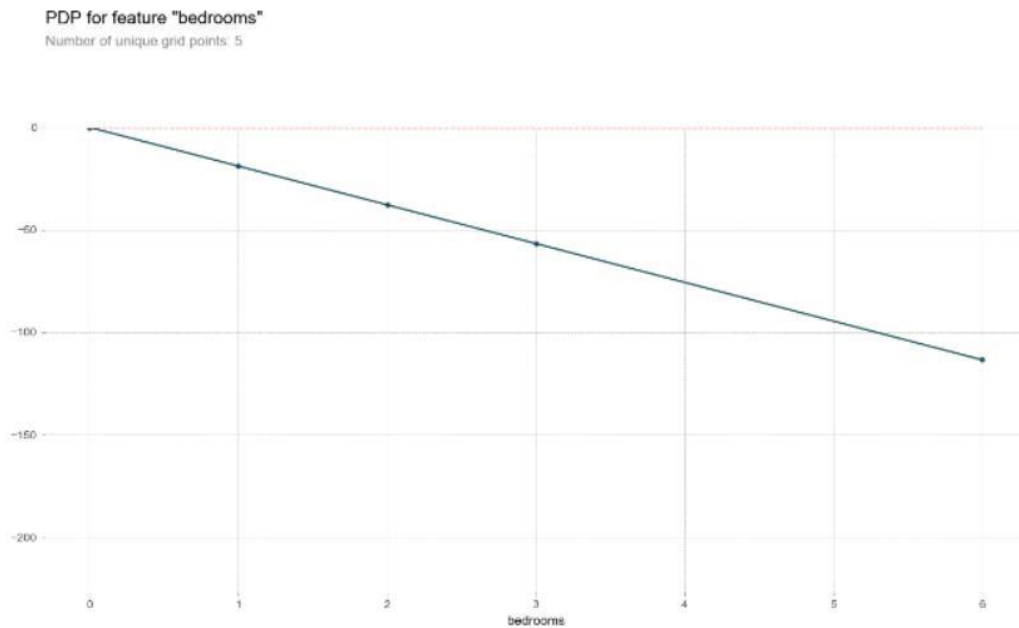
Number of unique grid points: 10



### PDP for feature "bathrooms"

Number of unique grid points: 3





## Conclusion:

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