Price Prediction

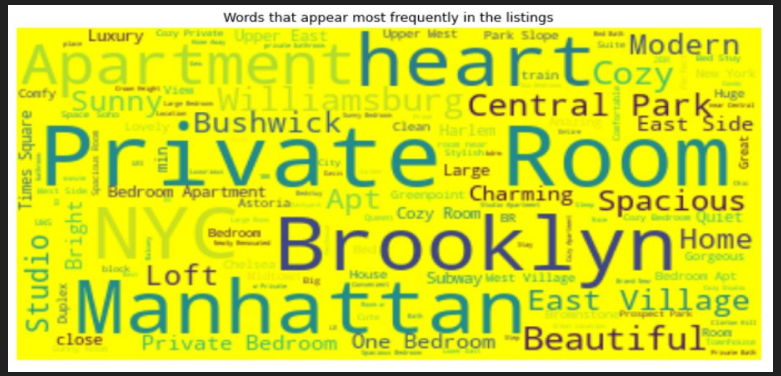
Predicting AirBnB rental prices using machine learning.

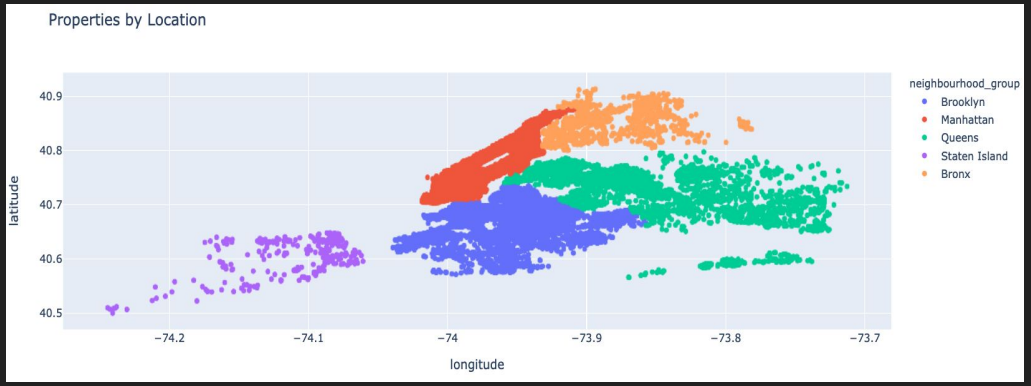
By Ishan Dahal

Motivation

- AirBnB is a great business model which allows everyday homeowners to rent out their homes for a fee
- Before COVID pandemic the AirBnB rental marker was thriving.
- It would be interesting to explore the data regarding rental properties
- Can we learn the relationship from features to the prices?

Dataset and Features





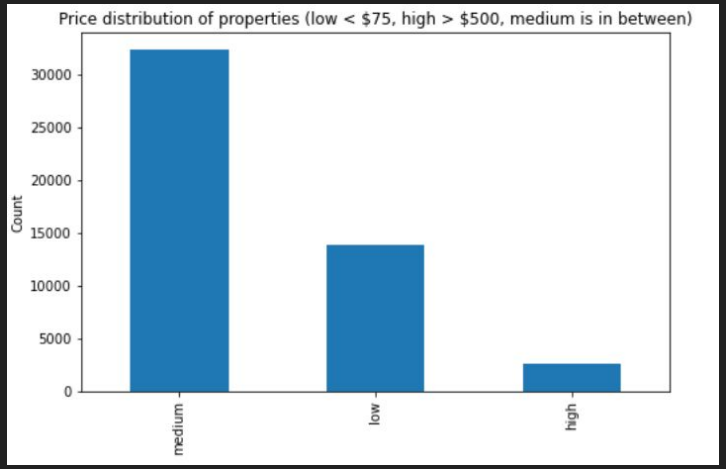
Dataset consisted of rental properties from different areas of NY city.

Target variable is price in dollars.

No significant relevant relationships betweens features were prevalent during feature analysis.

Manhattan and Brooklyn have highest density of properties.

Dataset was downloaded from kaggle.



Methods`

- The dataset is structured, so decided to use traditional machine learning algorithms.
- Feature analysis was performed to inspect distributions and relationships.
- Features were both continuous and discrete (categorical).
- Categorical variable were transformed using “one hot encoding” and continuous values were normalized.

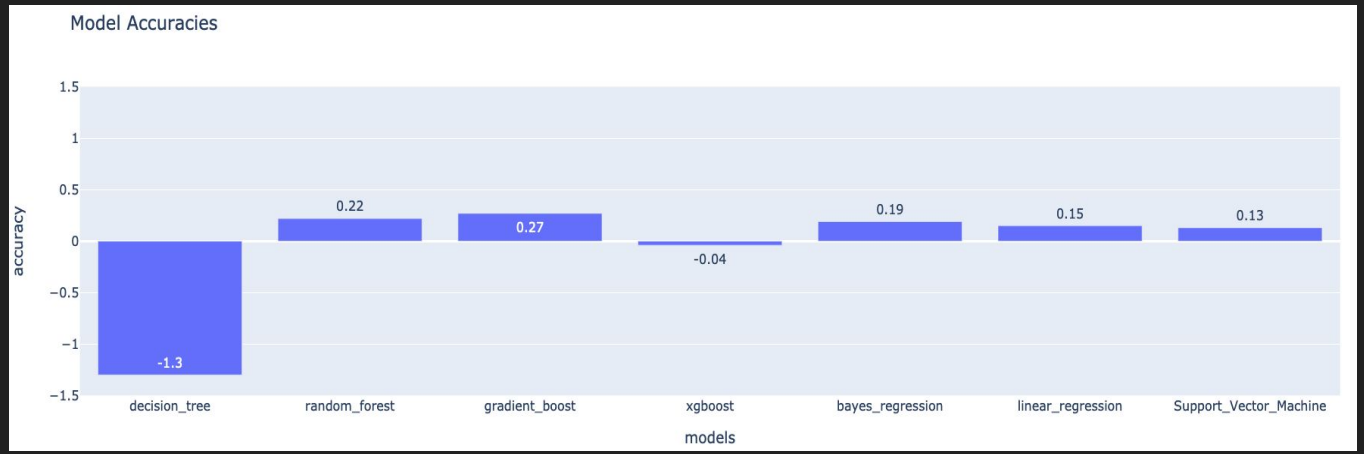
Methods

- Dataset was split into training, test and validation set.
- Test set was used to get an unbiased estimate of the model.
- Algorithms used: Bayesian Regressor, Decision tree, Random Forest, Gradient Boosting (sklearn), XGBoost, Linear regression and Support Vector Machines

Baseline and Modeling

- Decision tree was used to create a baseline model
- One hold out set method with validation set was used to estimate the model performance
- Grid search was used to calculate optimal parameters for gradient boosting
- Time to complete grid search was 10+ hours
- Single model for XGBoost took ~30 minutes to train and accuracy estimate varied

Result and Discussion



	accuracy_score
decision_tree	-1.30
random_forest	0.22
gradient_boost	0.27
xgboost	-0.04
bayes_regression	0.19
linear_regression	0.15
Support_Vector_Machine	0.13

- R2 score was used as a scoring metric where score of 1 is perfect performance and 0 is no learning.
- Gradient boosting from scikit learn was the best performer in predicting the price.
- Decision tree was the worst performer.

Conclusion and Future Steps

- Trying different transformations to the features with continuous values.
- Looking at deep learning model (fully connected network) to see if we can improve the accuracy.
- Use cross validation to get a better estimate of the model performance.
- Gathering more features especially pertaining to the dates of rental could be helpful.