Airbnb Listing, Metrics in NYC, NY, USA (2019); Price Prediction using Data Robotics

March 2022

ABUSOMWAN AMADEVBORHO
IDOWU JAMES
AJULO JOSEPH

DataRobot

TUS Midlands Midwest

Contents

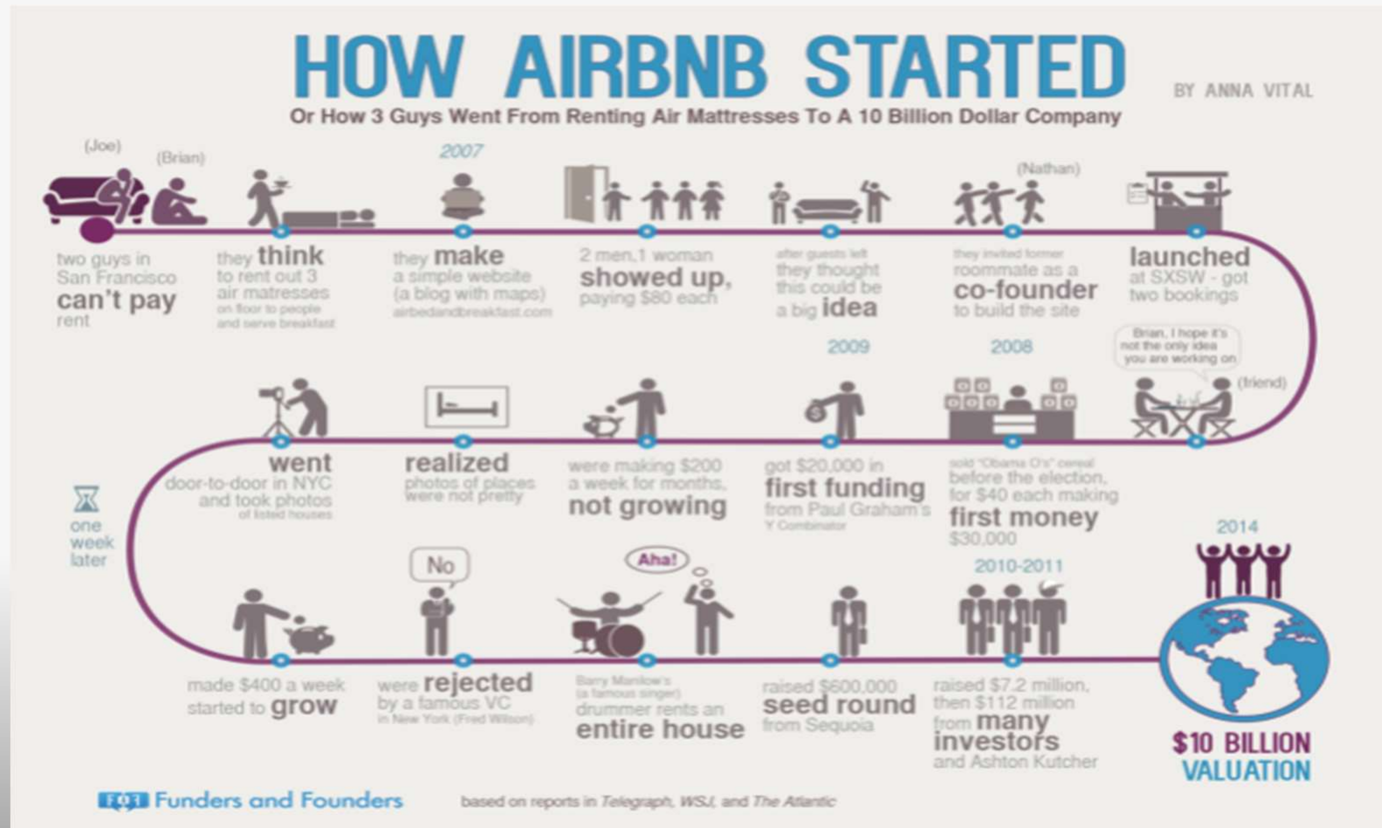
Introduction

We are going to be working on a large time series dataset. This dataset is a csv file which looks at and describes the AirBnB listing activities and metrics in New York city for 2019, containing 16 columns and 48,896 rows. The dataset consists of various attributes such as id, name, host id, host name, neighbourhood, neighbourhood group, latitude longitude, room type, price, minimum night, reviews, availability etc which we would be exploring during this project.

Dataset Source: AirBnB New York City 2019

AirBnB History

Airbnb is an online American company which was established in 2007 specializes in online housing accommodation services such as for lodging, primarily homestays like vacations, rentals and tourism activities.



Motivation of Study

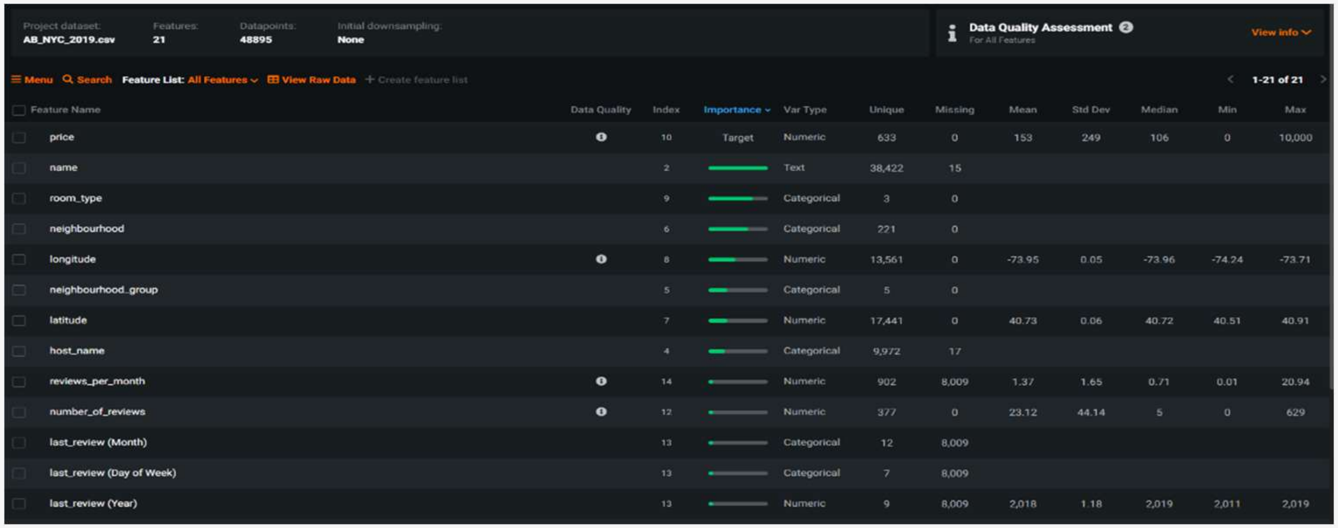
This dataset shows cumulation of information that we can find out more about such as hosts, geographical availability, necessary metrics to make predictions and draw conclusions.

With the suitable information there were various discussions that arose ranging from different the hosts and their areas, why traffic was also different in these areas, we could discover why some host are the busiest and why.

Thus, the ultimate question was created, that can we use this information to create a model which will give a suitable price prediction for future bookings.

Summary Statistics

Summary Statistics



Project dataset: AB_NYC_2019.csv | Features: 21 | Datapoints: 48895 | Initial downsampling: None

Data Quality Assessment For All Features | View info

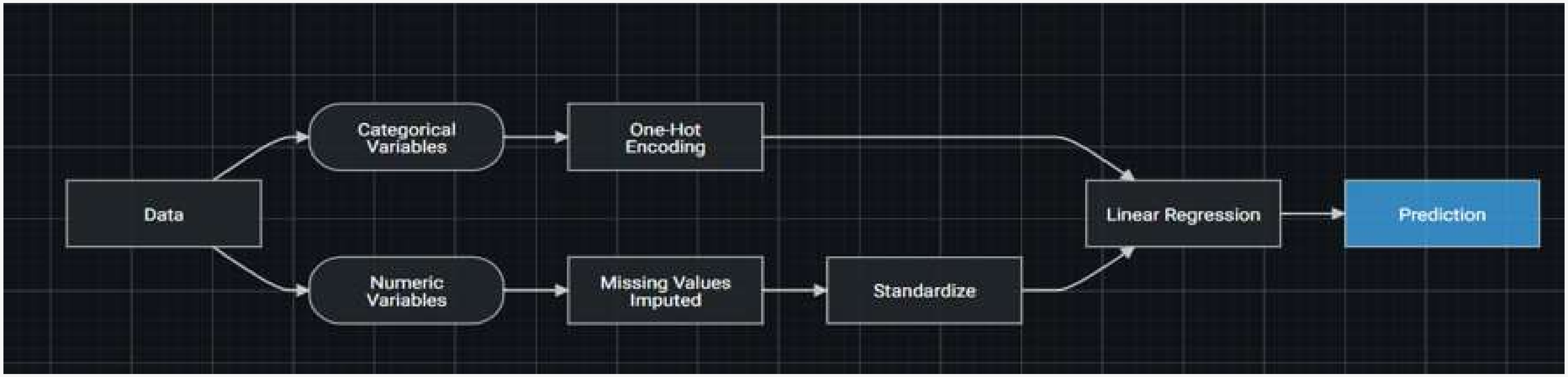
Menu | Search | Feature List: All Features | View Raw Data | Create feature list | 1-21 of 21

Feature Name	Data Quality	Index	Importance	Var Type	Unique	Missing	Mean	Std Dev	Median	Min	Max
price		10	Target	Numeric	633	0	153	249	106	0	10,000
name		2		Text	38,422	15					
room_type		9		Categorical	3	0					
neighbourhood		6		Categorical	221	0					
longitude		8		Numeric	13,561	0	-73.95	0.05	-73.96	-74.24	-73.71
neighbourhood_group		5		Categorical	5	0					
latitude		7		Numeric	17,441	0	40.73	0.06	40.72	40.51	40.91
host_name		4		Categorical	9,972	17					
reviews_per_month		14		Numeric	902	8,009	1.37	1.65	0.71	0.01	20.94
number_of_reviews		12		Numeric	377	0	23.12	44.14	5	0	629
last_review (Month)		13		Categorical	12	8,009					
last_review (Day of Week)		13		Categorical	7	8,009					
last_review (Year)		13		Numeric	9	8,009	2,018	1.18	2,019	2,011	2,019

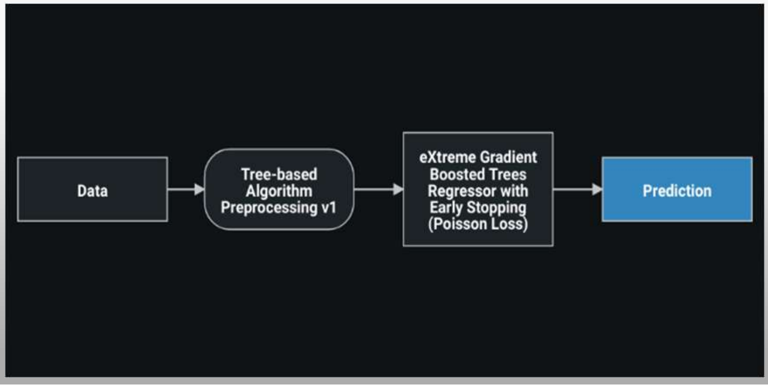
Research Methodology/Modelling

Blueprints for the models after training of the dataset

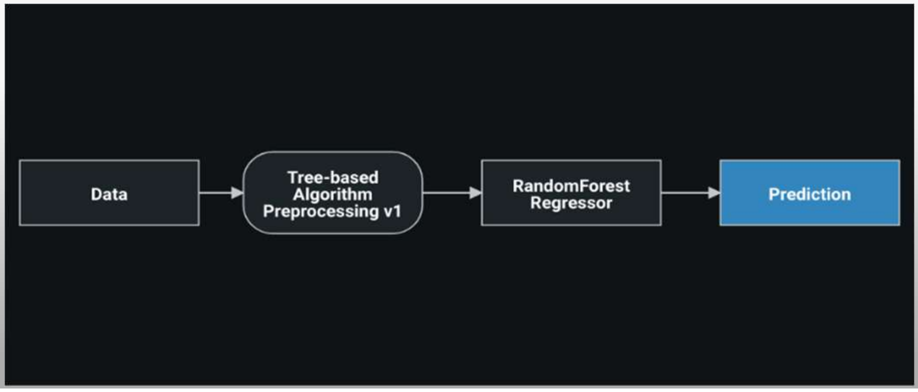
Linear Regression



eXtreme Gradient Boosted Trees Regressor with Early Stopping (Poisson Loss)

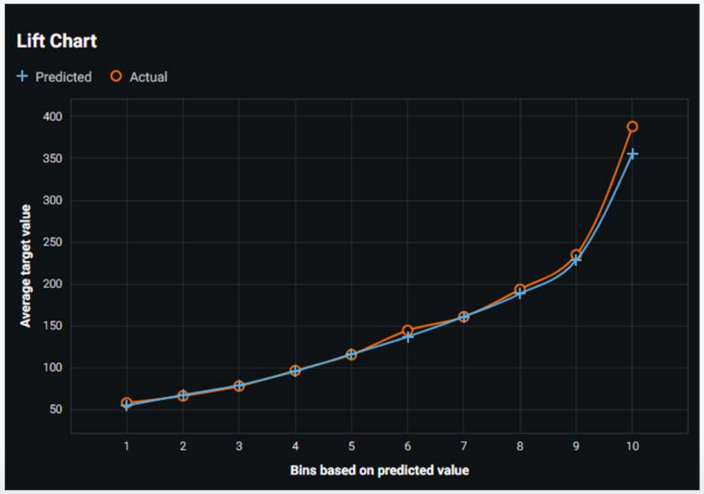


RandomForest Regressor



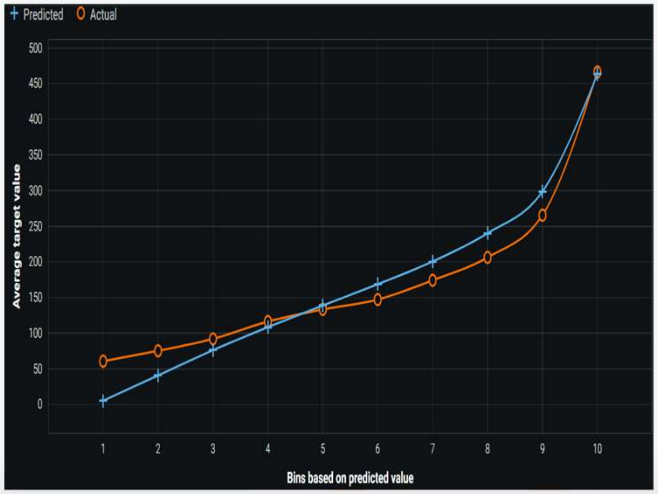
Lift Chart

eXtreme Gradient Boosted Trees Regressor with Early Stopping (Poisson Loss)



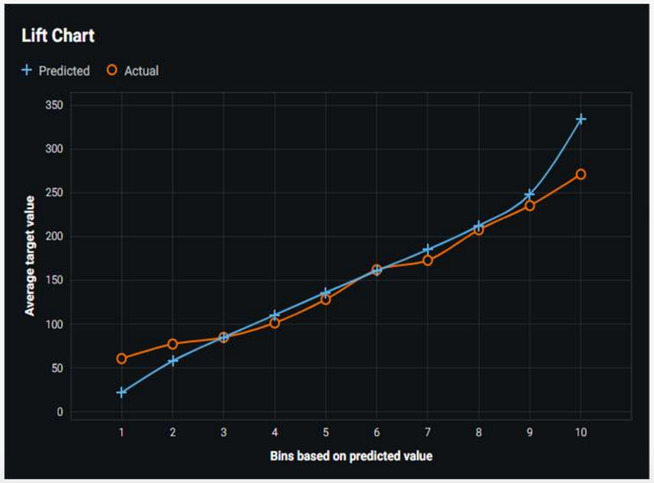
Prediction Time = 0.7663s

Random Forest Regressor



Prediction Time = 0.8726s

Linear Regression

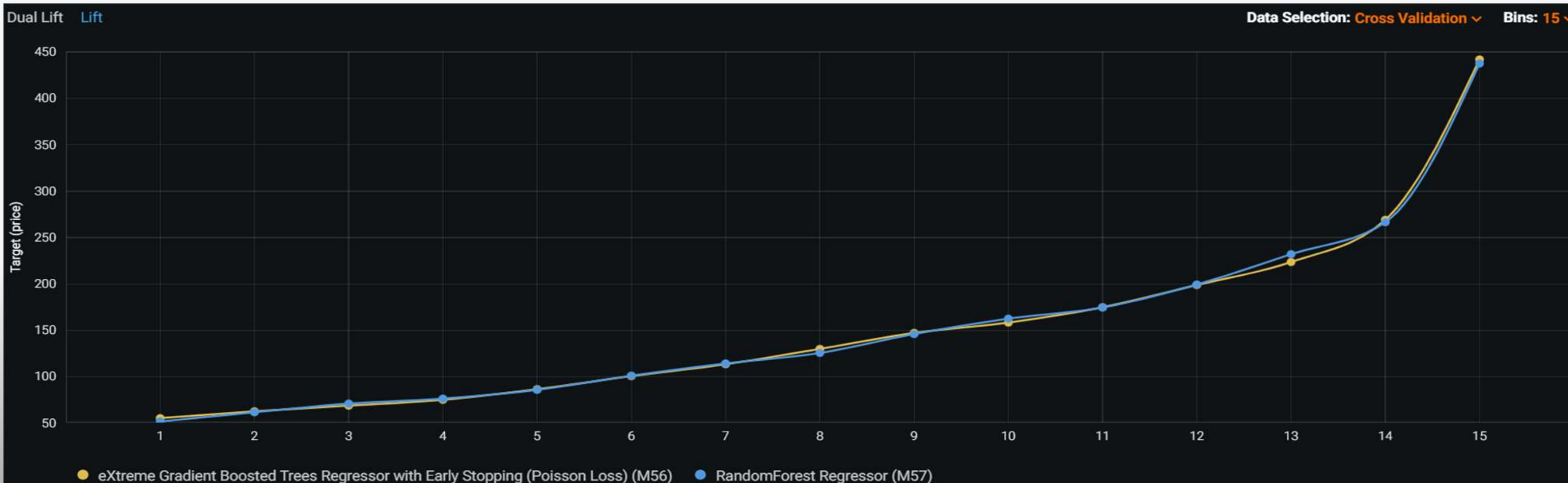


Prediction Time = 0.2765s

Model Comparison (Lift Trends)

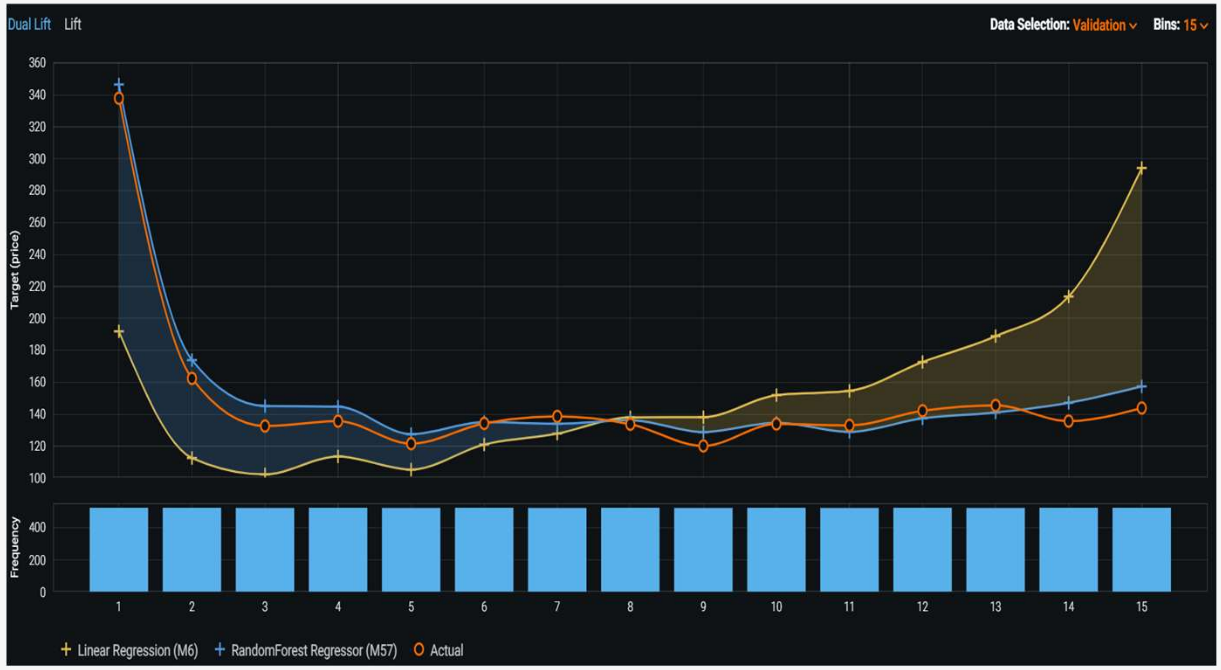
Comparing the extreme Gradient Boosted Tree Regressor and Random Forest Regressor here helps to identify a model that offers the highest business return or prediction. The Insights below shows a relatively strong model for comparison.

<table>
<tr><th>Model</th><th>RMSE (Validation)</th><th>RMSE (Cross Validation)</th><th>Gini Norm (Validation)</th><th>Gini Norm (Cross Validation)</th><th>Prediction Time (s)</th></tr>
<tr><td>eXtreme Gradient Boosted Trees Regressor with E...
M56 | BP78 | 64% Sample Size | Informative Features</td><td>157.5703</td><td>218.2339</td><td>0.7710</td><td>0.7773</td><td>7.25</td></tr>
<tr><td>RandomForest Regressor
M57 | BP66 | 64% Sample Size | Informative Features</td><td>158.2511</td><td>220.9422</td><td>0.7628</td><td>0.7777</td><td>11.57</td></tr>
</table>



Model Comparison (Dual Lift)

Comparing the Linear Regression Model and Random Forest Regressor with large difference in prediction time and sample size. This Insights below shows a relatively weak model for comparison.

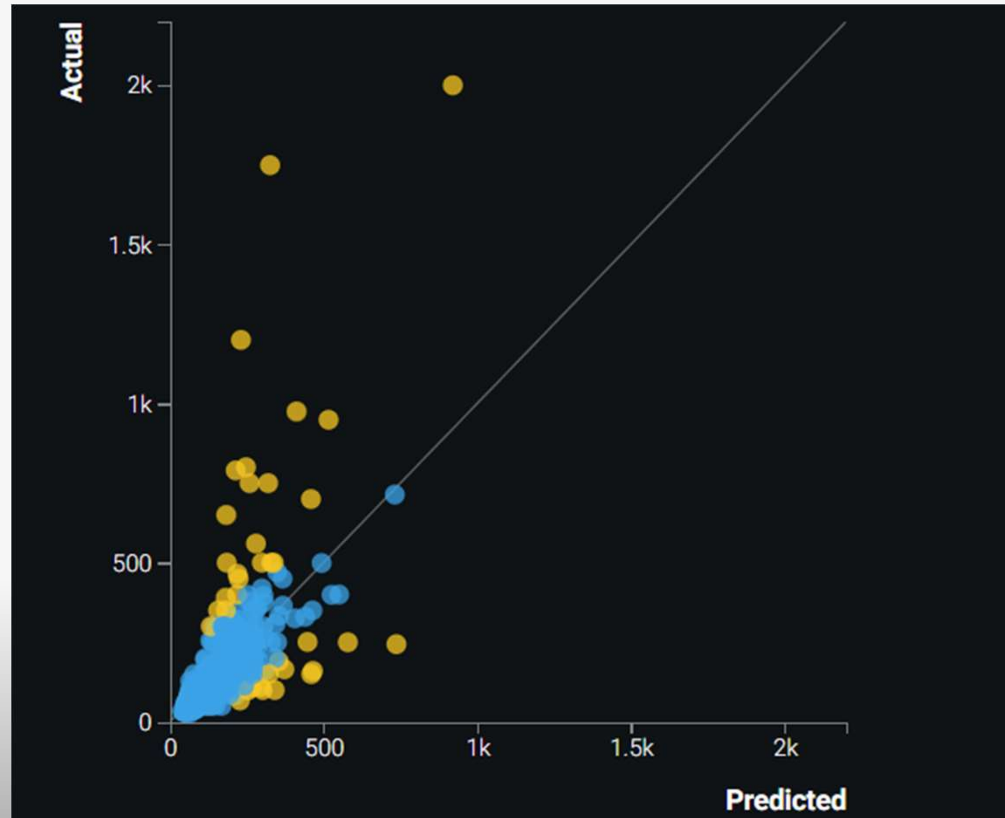


Model	RMSE (Validation)	RMSE (Cross Validation)	Gini Norm (Validation)	Gini Norm (Cross Validation)	Prediction Time (s)
Linear Regression M6 | BP51 | 16% Sample Size | Informative Features	175.5015	N/A	0.6395	N/A	3.54
RandomForest Regressor M57 | BP66 | 64% Sample Size | Informative Features	158.2511	220.9422	0.7628	0.7777	11.57

Predictive Performance and Regression Model Adopting the Random Forest Regressor

Accuracy Parameters

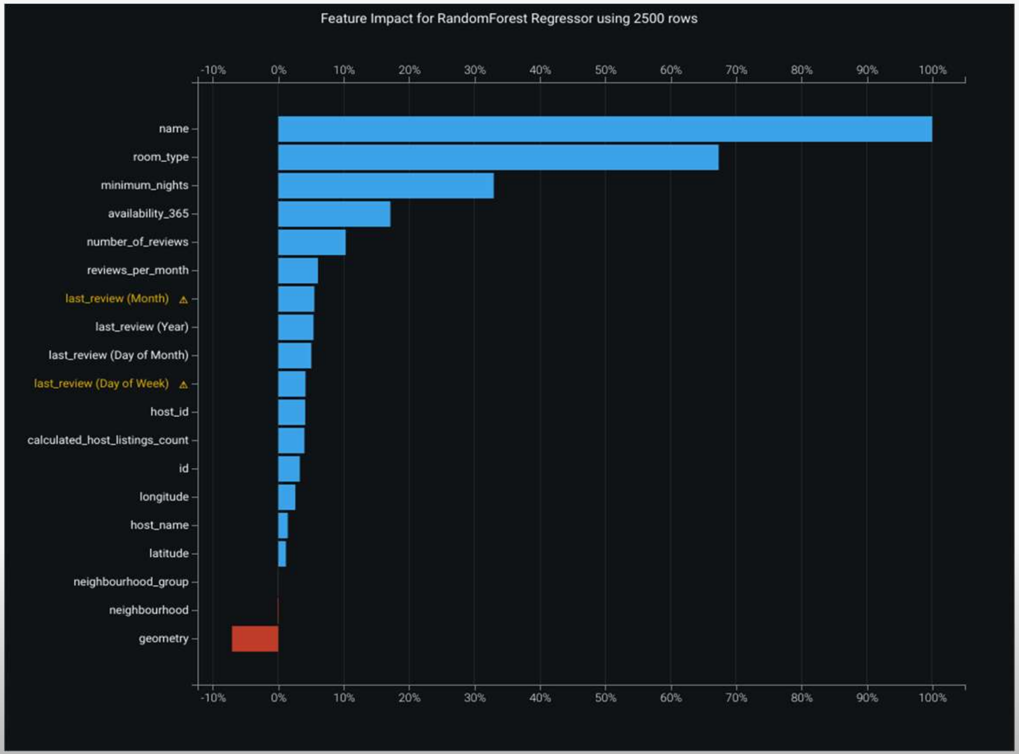
Residual mean: -3.8051 Coefficient of determination (r^2): 0.3325 Standard Deviation: 155.3118

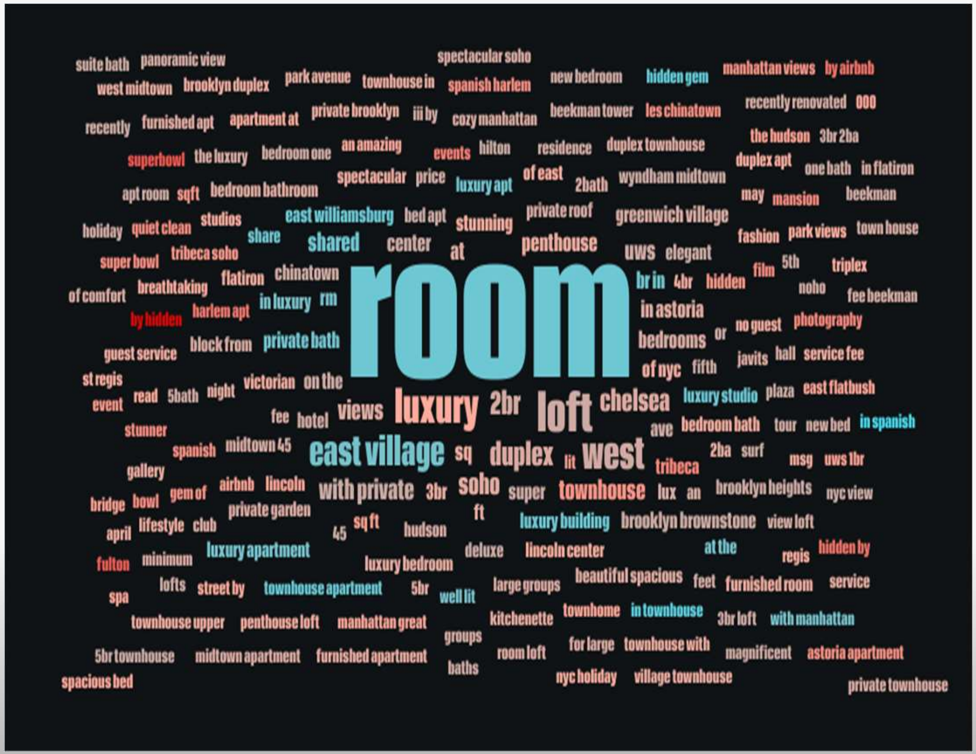


Predictive Performance and Validity of the Regression Model.

Insights

From the insight option we can tell the importance of each feature. The AirBnB name has a greater impact on the price. The word cloud shows how different texts from the name feature are associated with the predicted price.





Findings

- Theater district have the highest price for an Airbnb and area Bedford is the busiest neighborhood because of its relatively low price
- The entire home/ apartment commands the highest price
- Guest who spent 30 night least are much
- The name of the Airbnb play a very key role in modeling process

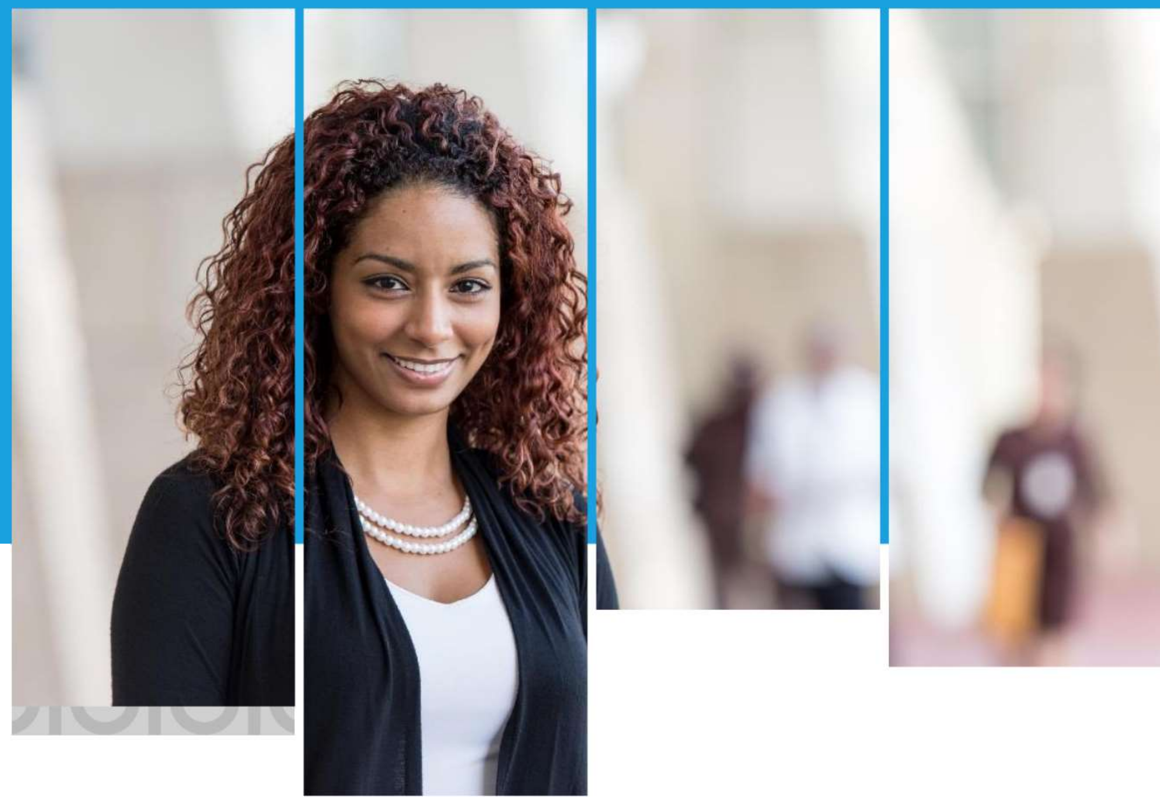
CONCLUSION AND RECOMMENDATION

DataRobot was very useful to our analysis providing EDA, Future Engineering, Data Cleaning, Model Training, piper parameter tuning and as well model deployment in seconds without writing any code.

Furthermore, we could try to perform as many analysis for value predictions via our models from data robot, However it certainly wouldn't match the reality, as there are few qualifying attributes for a property in New York and the location of the property and its price are not always guarantee of occupancy return.

Finally, DataRobot which we found as too automated. We would recommend its users to have more control of its functionalities happening behind the scene.

Thank you



ABUSOMWAN AMADEVBORHO
IDOWU JAMES
AJULO JOSEPH