Price i$ Right

Airbnb - Listings

Subhash Jaldu

Motivation

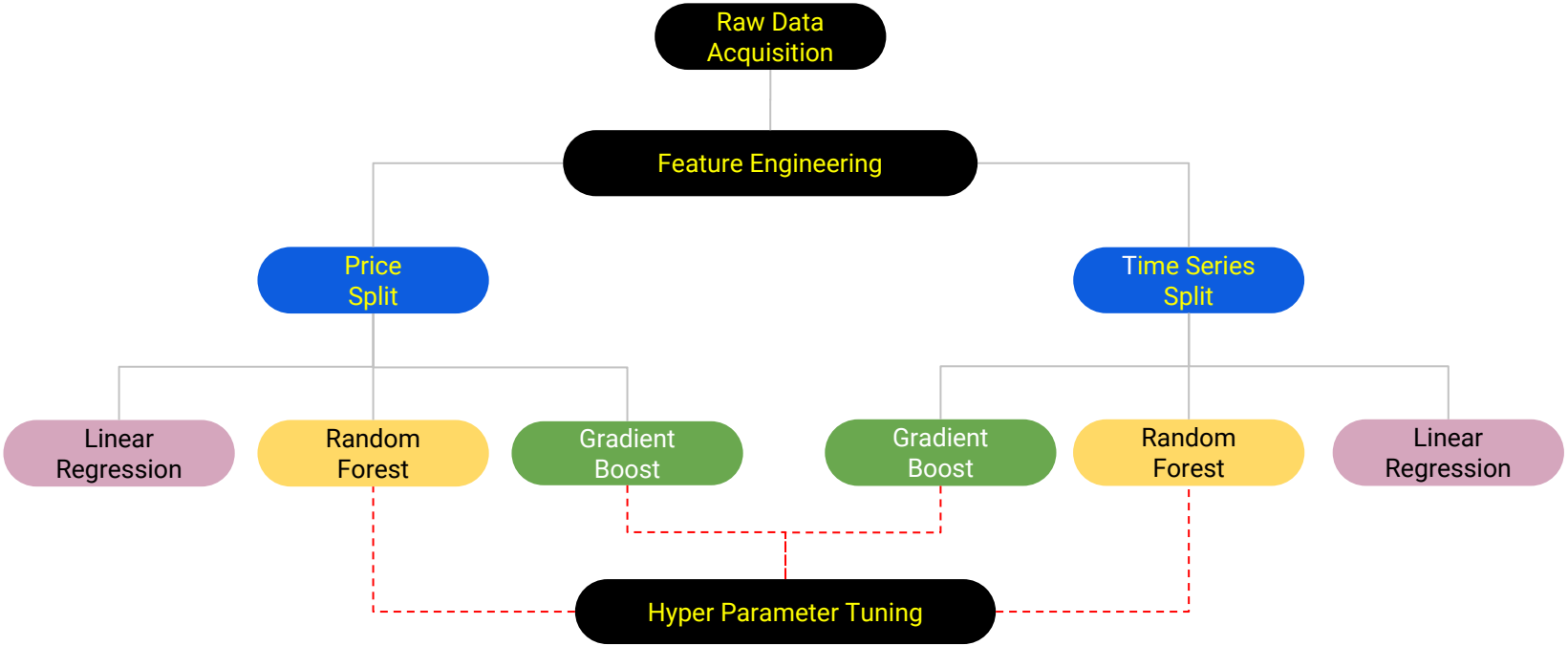
- Help the new host to list his property at FAIR price
- Help new hosts to improve occupancy rate by FAIR pricing
- Need a model to predict the price of the UNIQUE listing
    - Features
    - Specalities
    - Surge
    - Seasonality

Dataset

S.No.	Data	Size	Age	Data Size
1	Calendar	1.8mil x 4		8 million
2	Listings	5000 x 96		~ 500K
3	Neighborhood	37 x 2		
4	Reviews	250K x 6	9 years	1.5 mil
5	Geolocation			SF, Bayarea, LA County

Rich Data with 96 features 5k listings = 500K data points/county

Flow


Raw Data Acquisition
Feature Engineering
Price Split
Time Series Split
Linear Regression
Random Forest
Gradient Boost
Gradient Boost
Random Forest
Linear Regression
Hyper Parameter Tuning

Feature Analysis

Data Cleanup
- Transform data
- Missing values
- Categorical

Relational Features

Category Condensation

Imputed sparse data sets
- Luxury
- Special Types

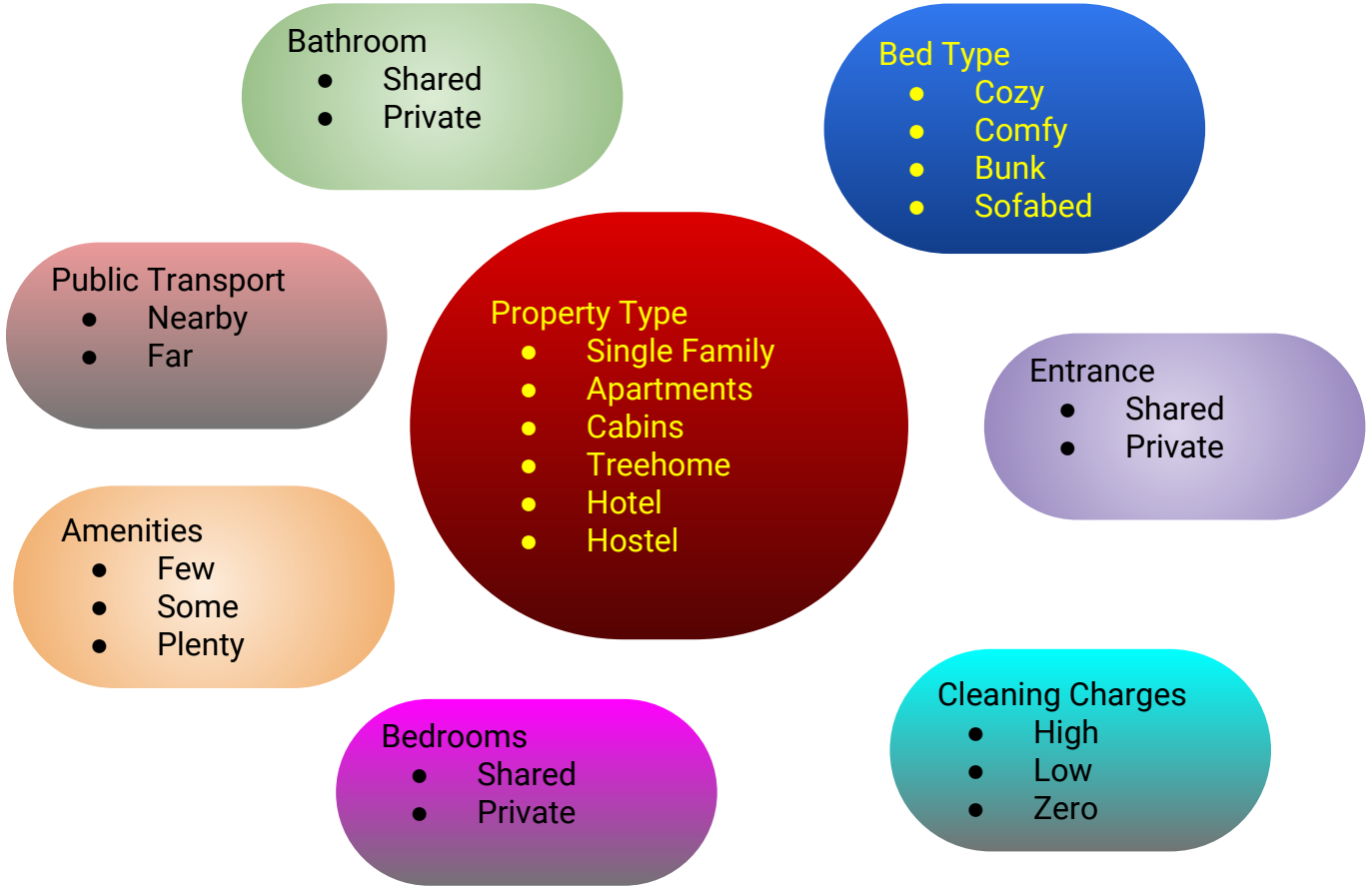
Feature Analysis

Data Cleanup
- Transform data
- Missing values
- Categorical

Relational Features

Category Condensation

Imputed sparse data sets
- Luxury
- Special Types



Model Ensemble

Models :

- Linear Regression
- Random Forest
- Gradient Boost

Data :

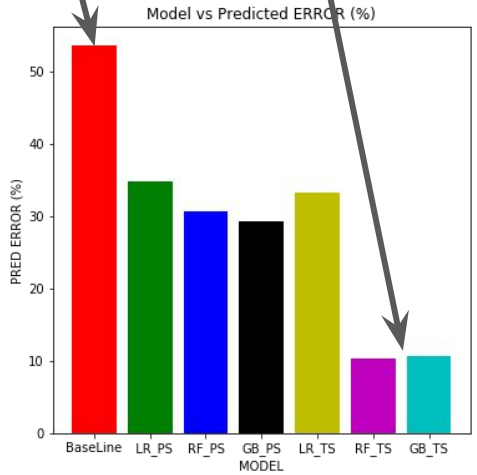
- Common Listings
- Time Series Split

Result Analysis

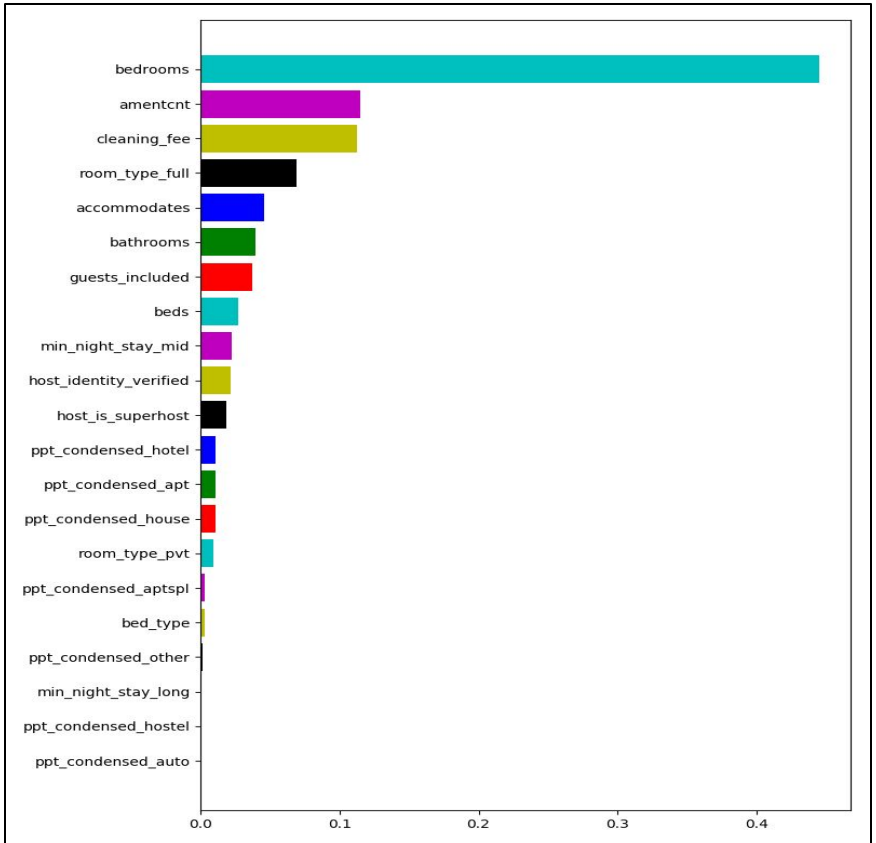
Metric		Linear Regression	Random Forest	Gradient Boost	Linear Regression	Random Forest	Gradient Boost
		Random Split			Time Series Split		
R^2		0.56	0.64	0.62	0.54	0.88	0.88
MSE		4155	3366	3578	4618	1187	1190
RMSE		64.46	58.02	59.82	67.96	34.46	34.50
RMSLE		0.38	0.33	0.33	0.37	0.19	0.20
PCT Error (MAPE)		34.86%	30.62%	29.20%	33.27%	10.37%	10.71%
Baseline Error		$77.42	$77.42	$77.42	$79.72	$79.72	$79.72
Predicted Error		$49.56	$43.98	$43.46	$50.37	$15.44	$15.90

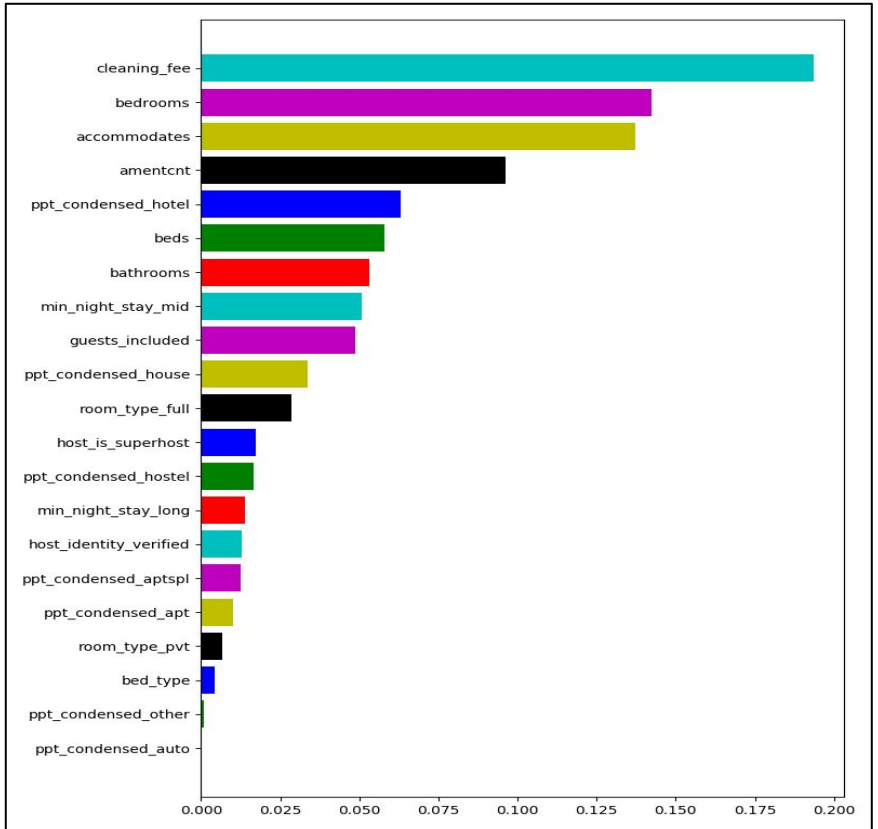
Baseline

Best FIT



Feature Importance (RF vs GB)





Result Analysis

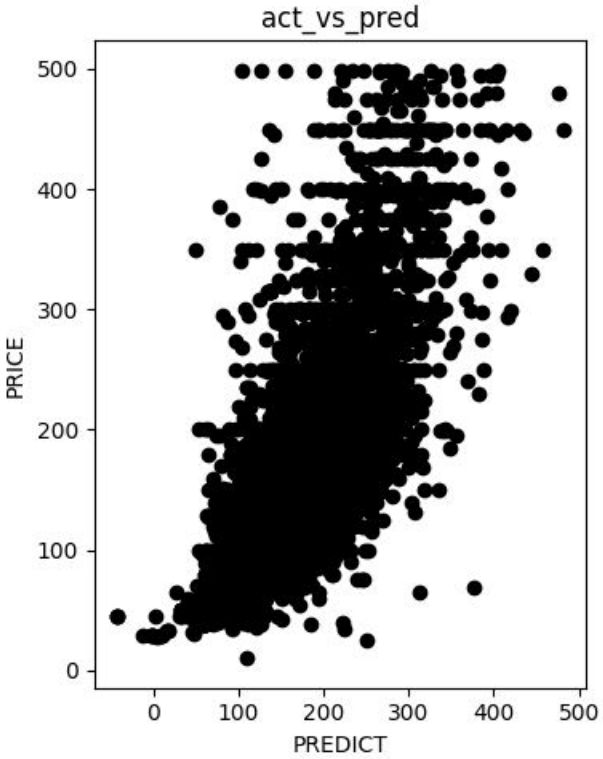
Prediction Metrics

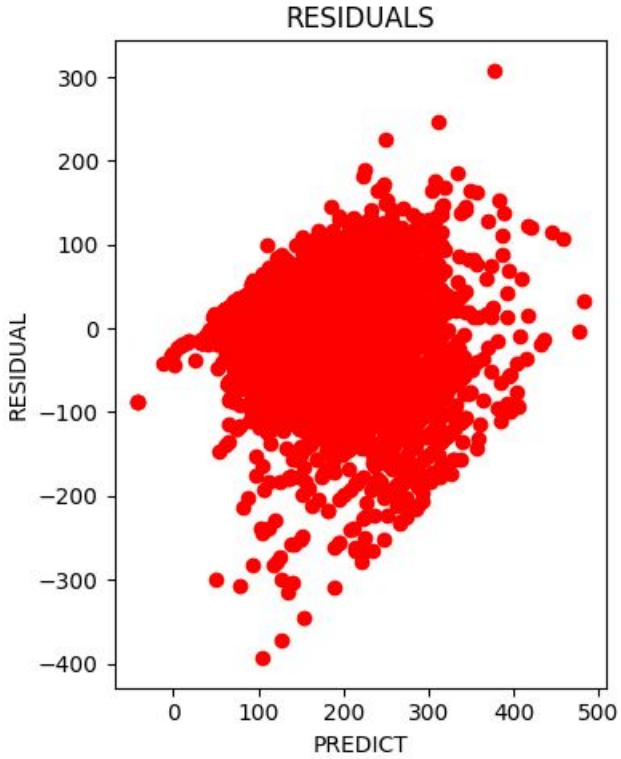
- R^2 : 0.8817
- PCT ERROR : 10%
- Baseline ERROR : $ 79.00
- Predicted ERROR : $ 15.90

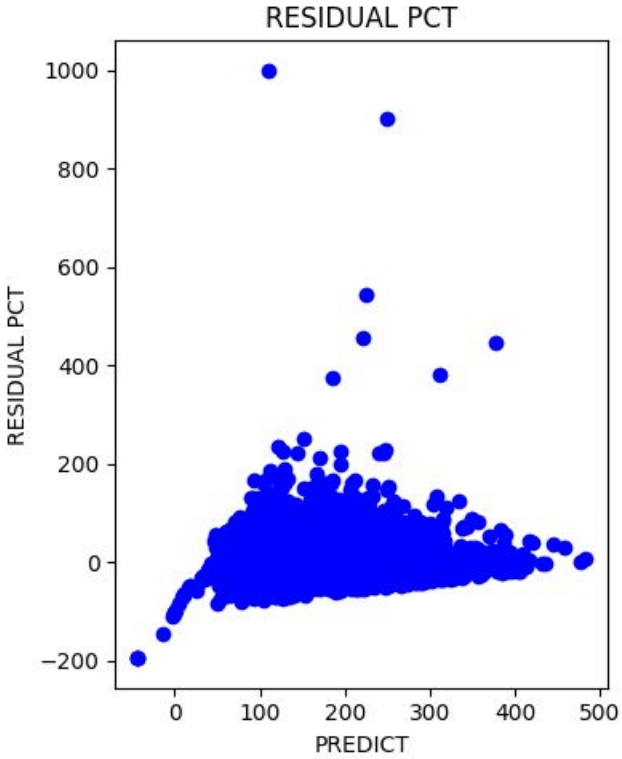
Ensemble :

- Random Forest and Gradient Boost
- Time Series

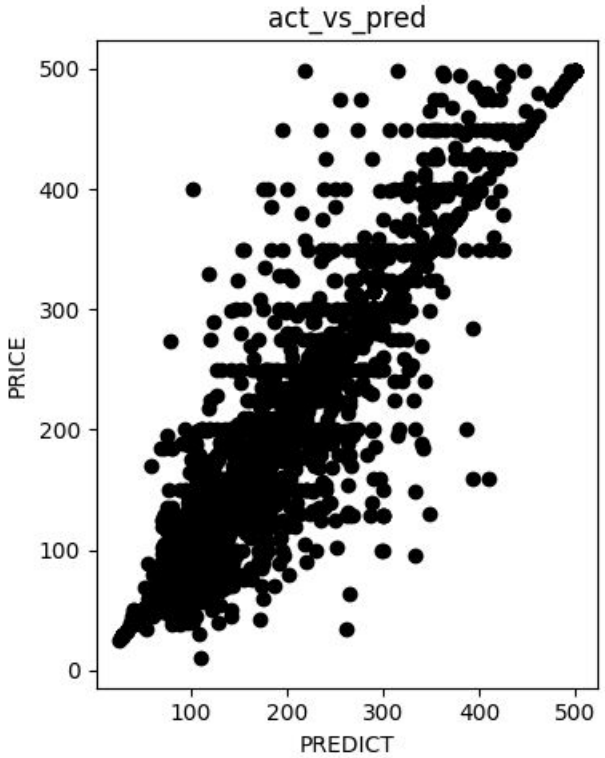
Model Metrics : Linear Regression

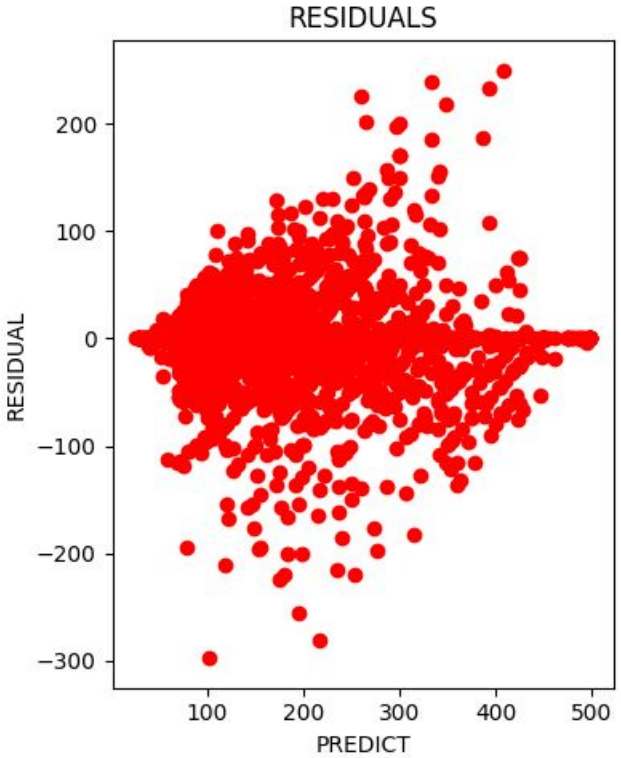


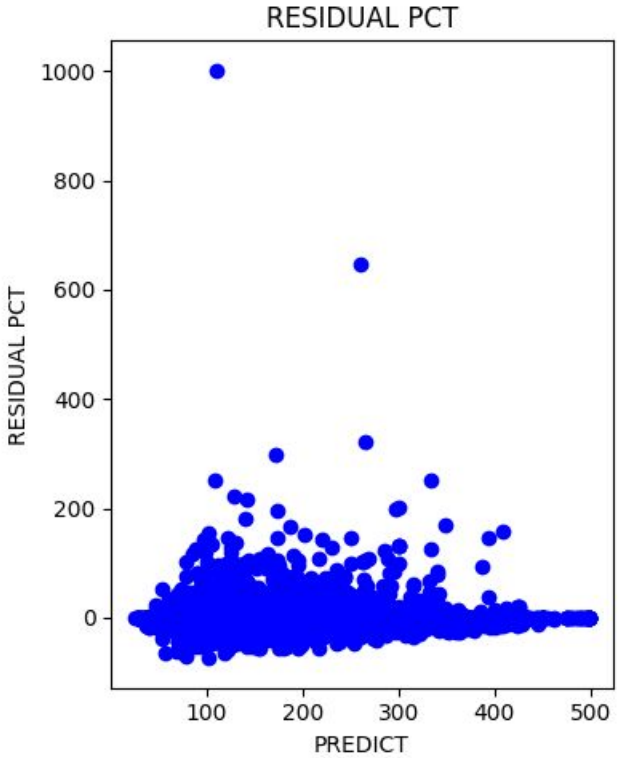




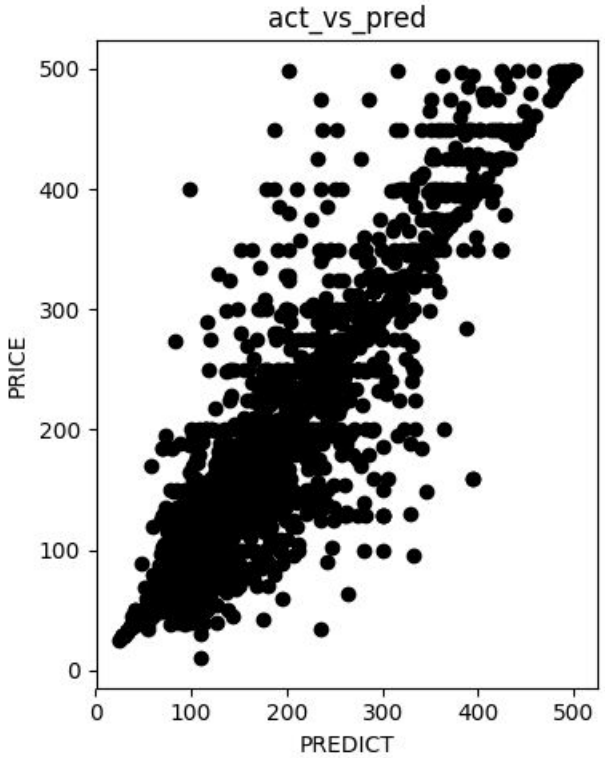
Model Metrics : Random Forest

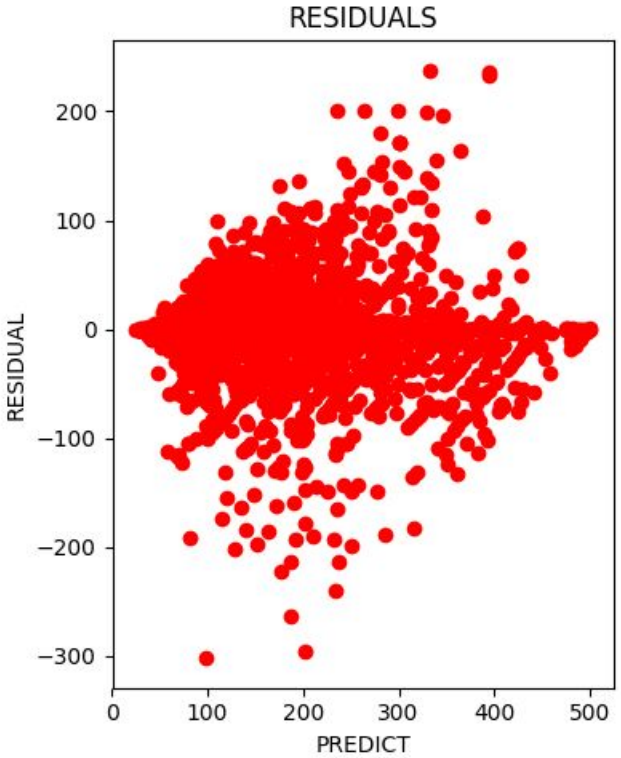


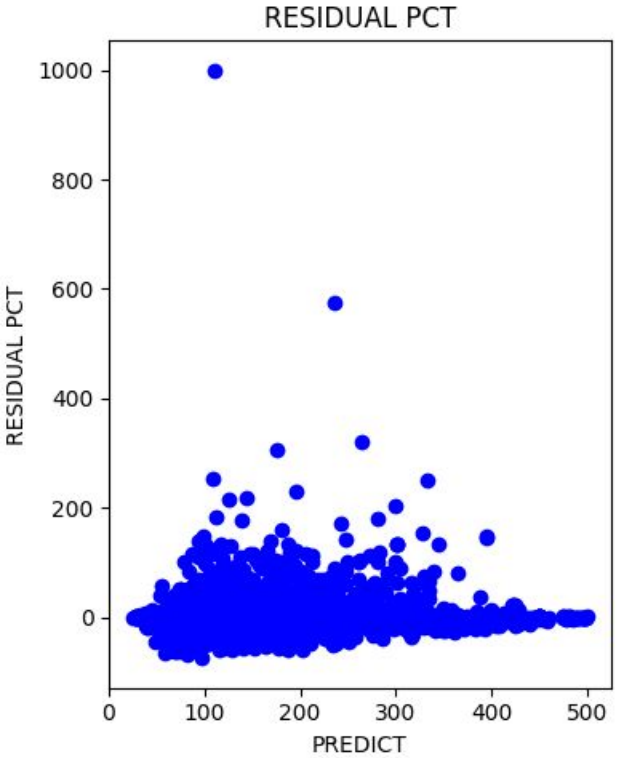




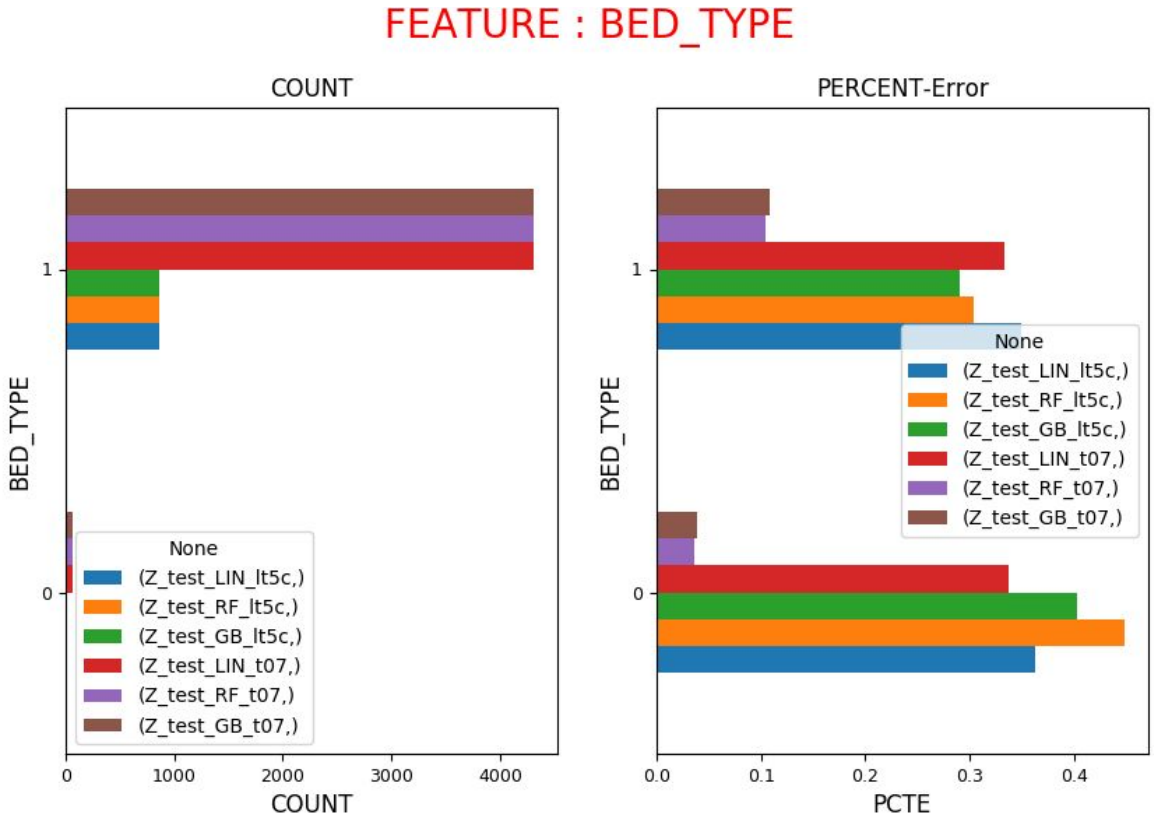
Model Metrics : Gradient Boost



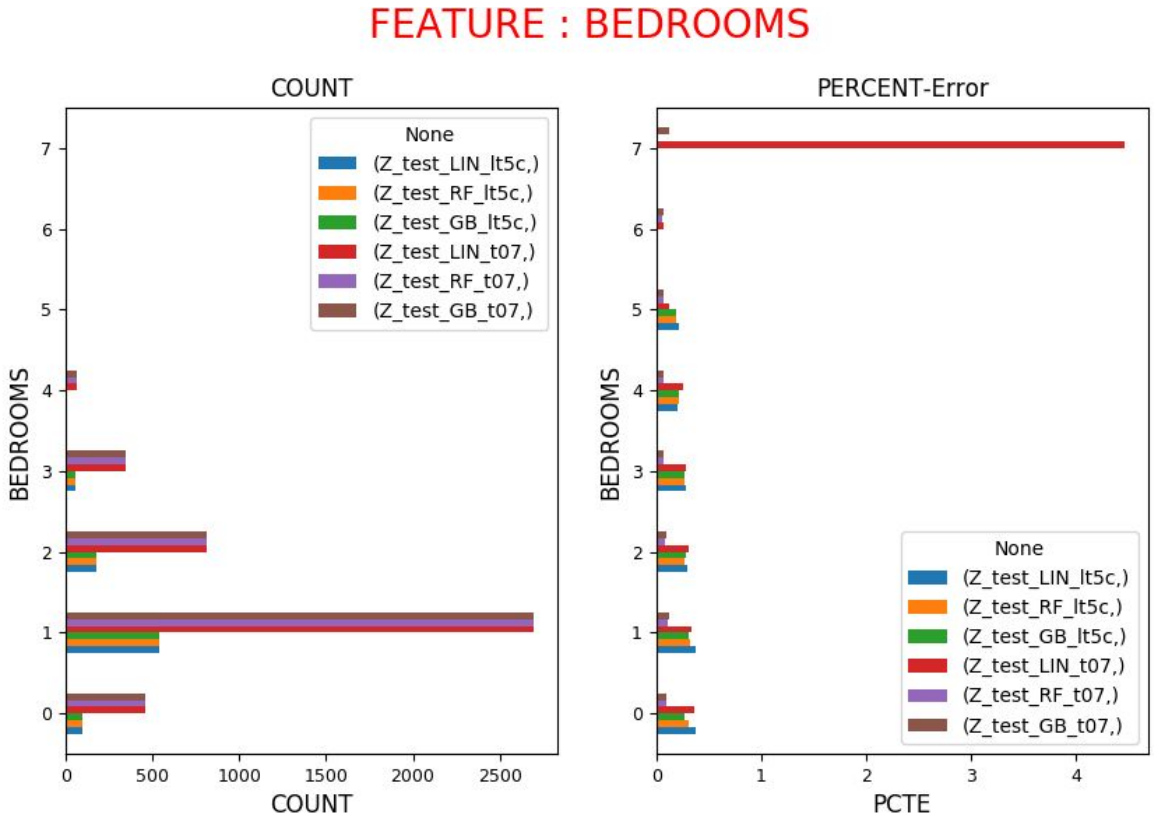




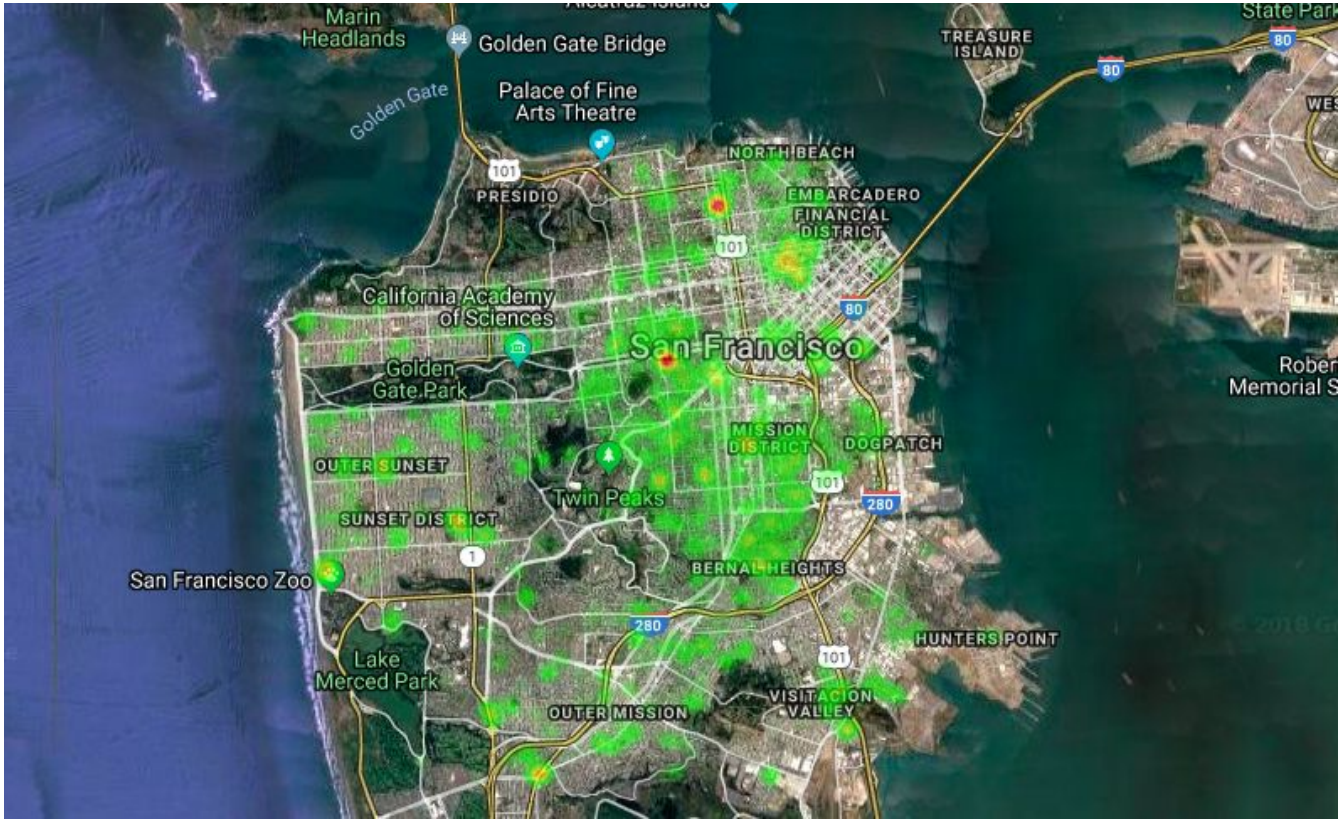
Model Performance Vs Bedtype



Model Performance Vs Bedrooms



Model Performance Vs Location



Future Improvements

- Add complete time series to include seasonal signal
- Analyze real time demand predict a surge for Host
- Analyze the reviews to provide a feature sentiment analysis to Host
- Update the model to support special cases

Thank you

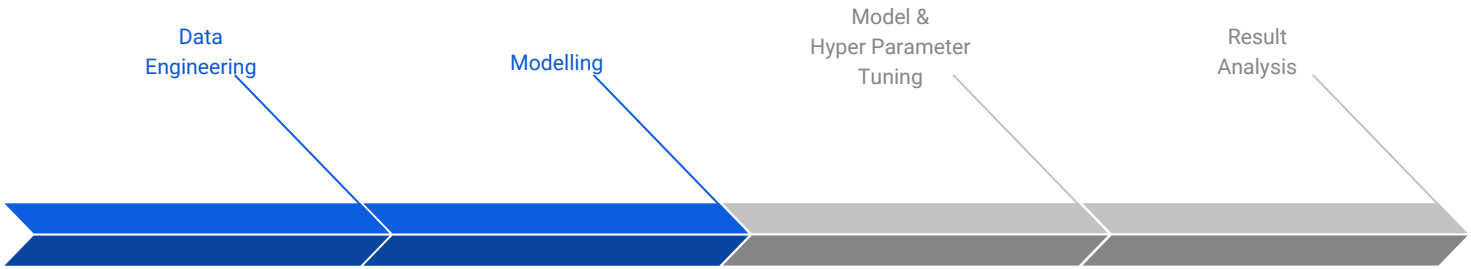
Questions?

Github : https://github.com/sgjcyp/capstone_price_is_right

Motivation

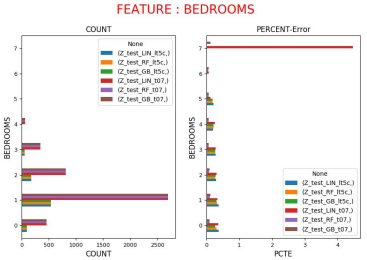
- Airbnb is a portal for homeowners to host the Property
- Typically Hosts review the current listings to decide the price
- Need a model to predict the price of the listing
    - Features
    - Uniqueness
    - Surge
    - Seasonality

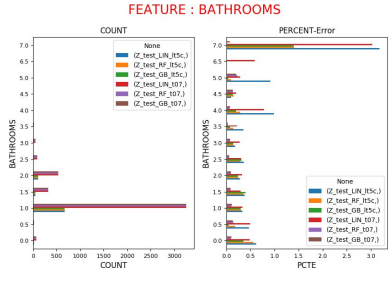
Process

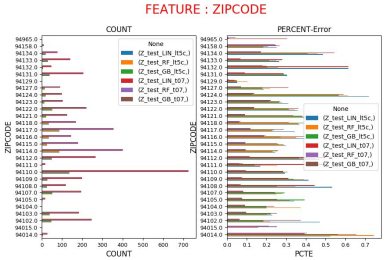


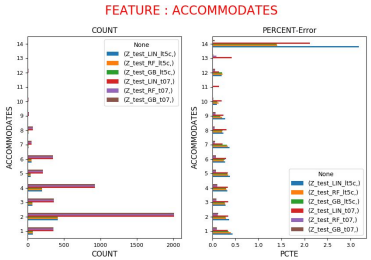
Feature Analysis	Regression Models	Measure Performance	Results vs Model
Identify Prime Features Create Relational Features	Linear Regression Random Forest Gradient Boost	Same Metrics to measure multiple models and datasets	Model Performance in Various Categories

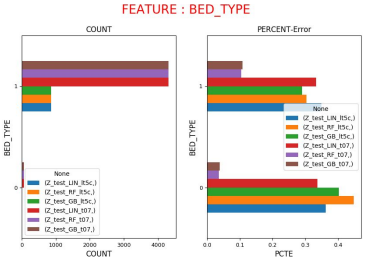
Model Performance Vs Features

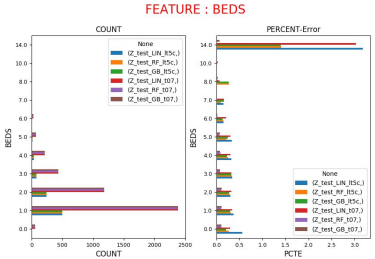




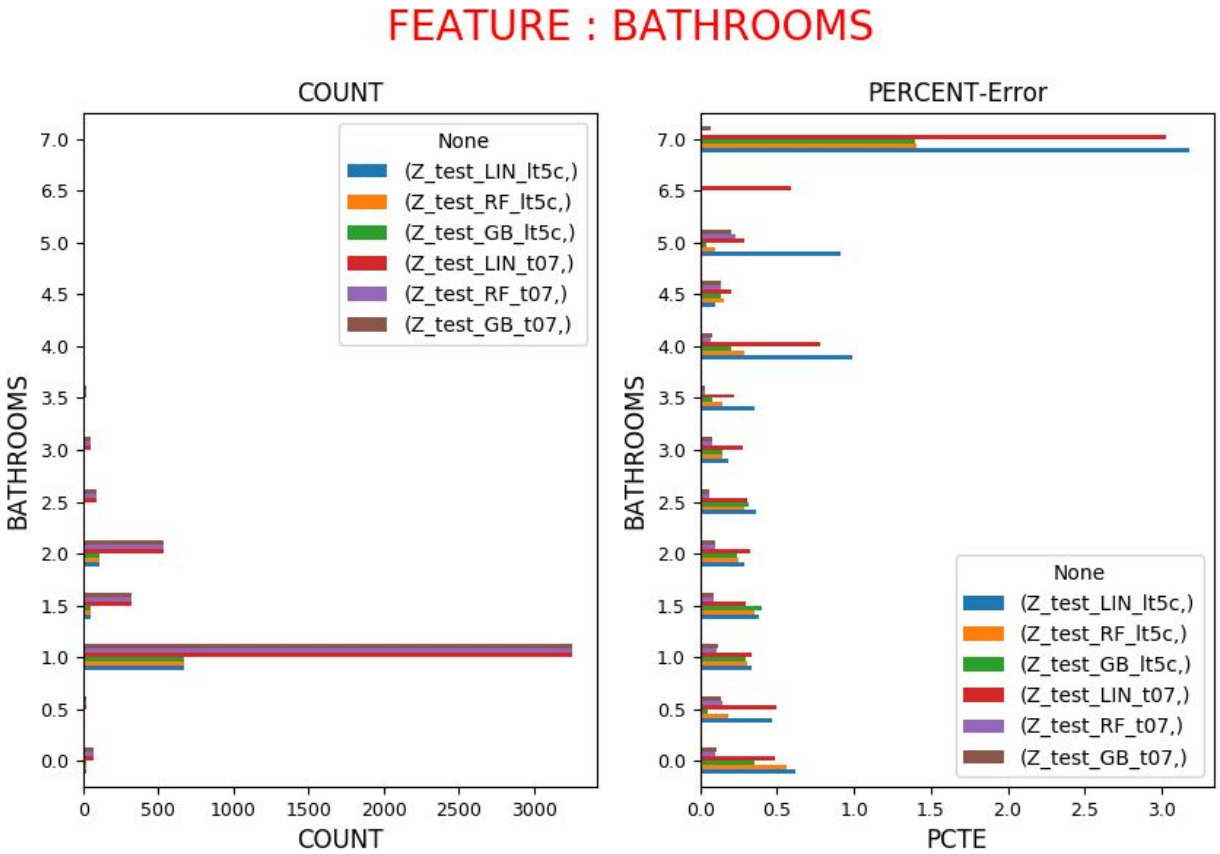




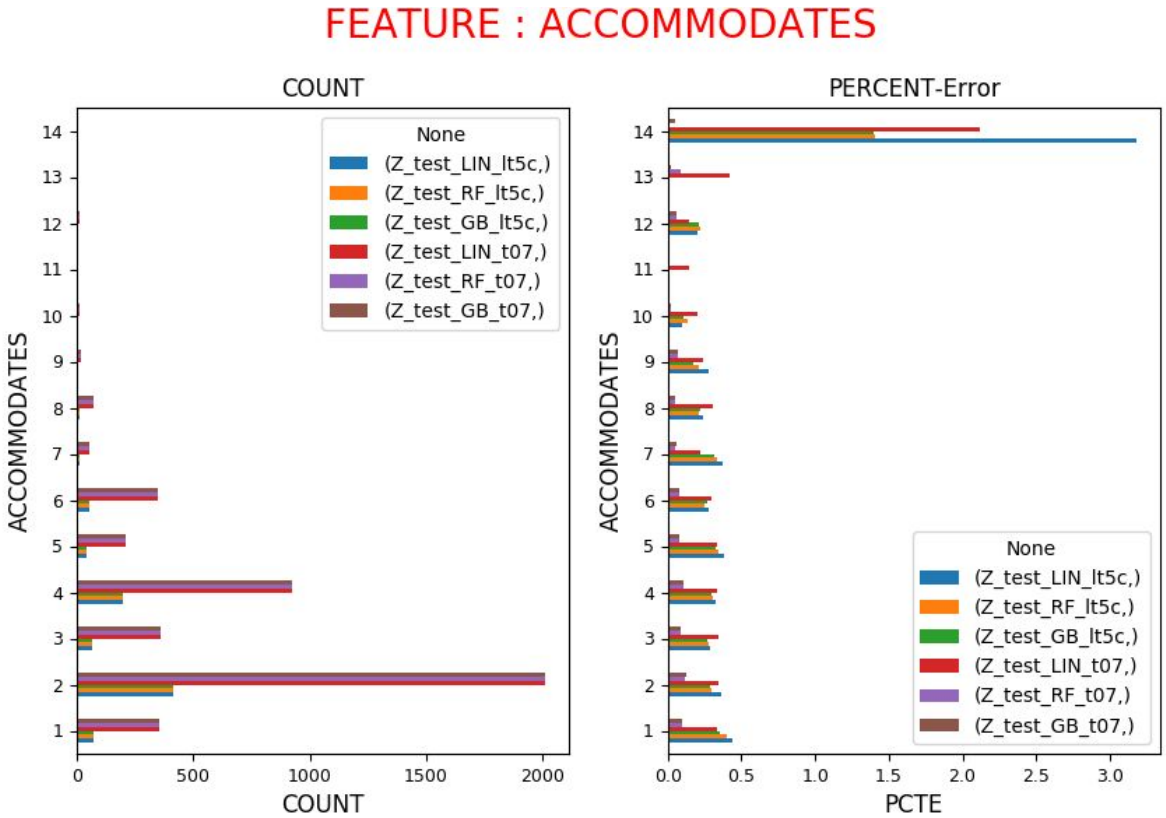




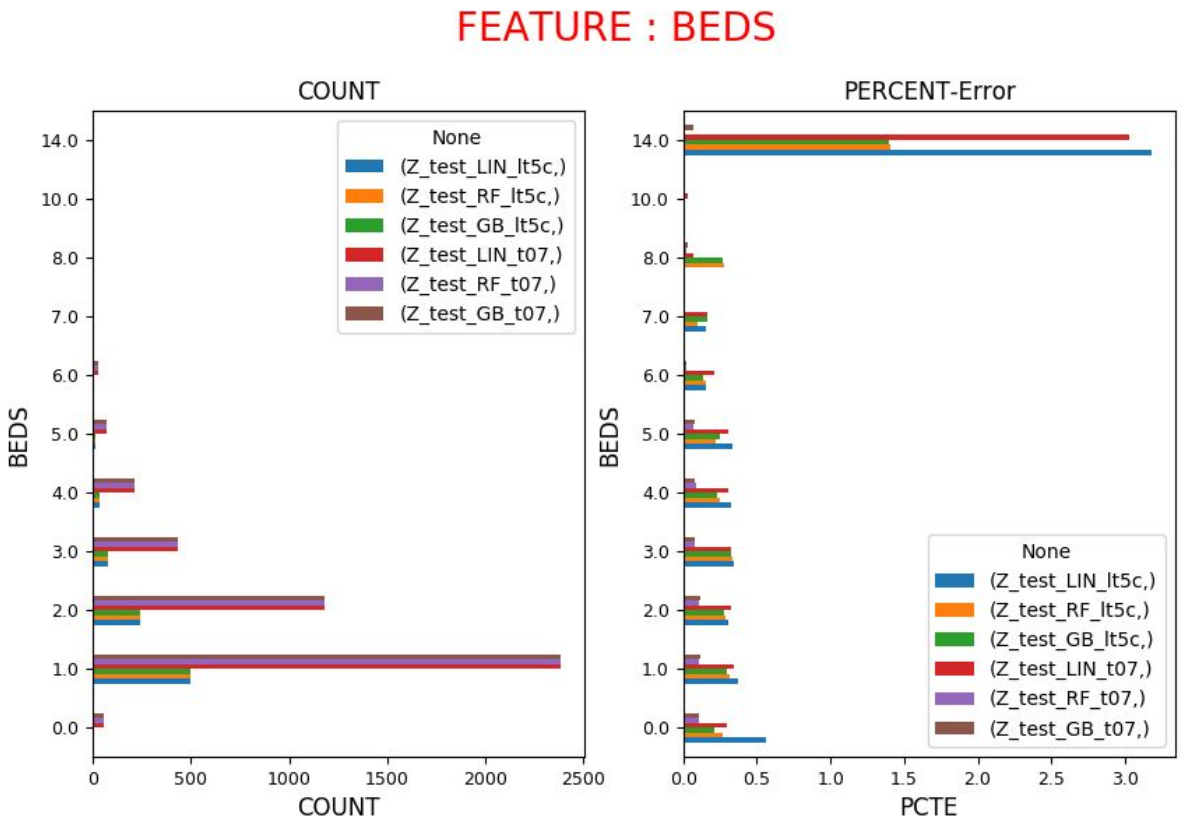
Model Performance Vs Bathrooms



Model Performance Vs Accommodates



Model Performance Vs Beds



Model Performance Vs Zipcode