FLATIRON ANALYTICS

AIRBNB PRICE ANALYSIS

ABOUT ME

JF ROBERTS

- **GIVING BACK!**
- **MENTAL HEALTH**
- **ENVIRONMENTAL CONSERVATION**

jfbr1283

jfwholehealth@gmail.com

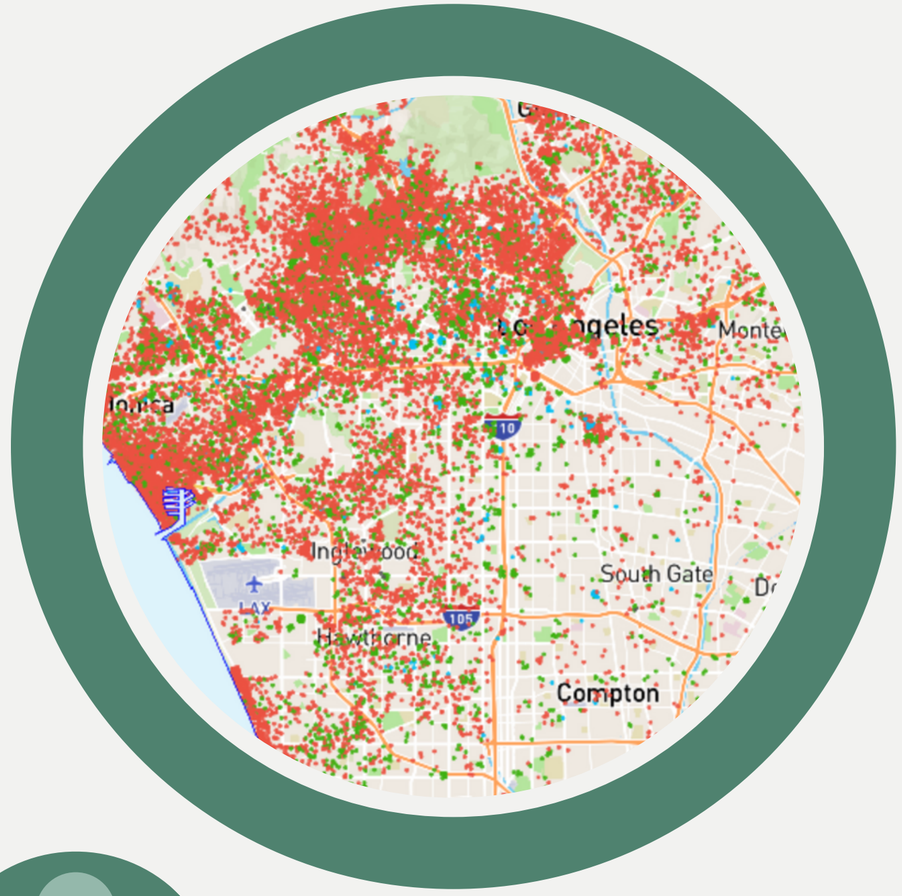
OVERVIEW

- BUSINESS PROBLEM
- DATA
- DATA PROCESSING
- PREDICTIVE MODELS: 3 GROUPS
- MODEL EVALUATION
- RECCOMENDATIONS & NEXT STEPS

BUSINESS PROBLEM

- CONTRACTED TO OPTIMIZE PRICE SETTING STRATEGY
- ANALYZE WRITTEN REVIEWS AS PRICE PREDICTORS
- RECOMMENDATIONS BASED ON MODEL PERFORMANCE

THE DATA



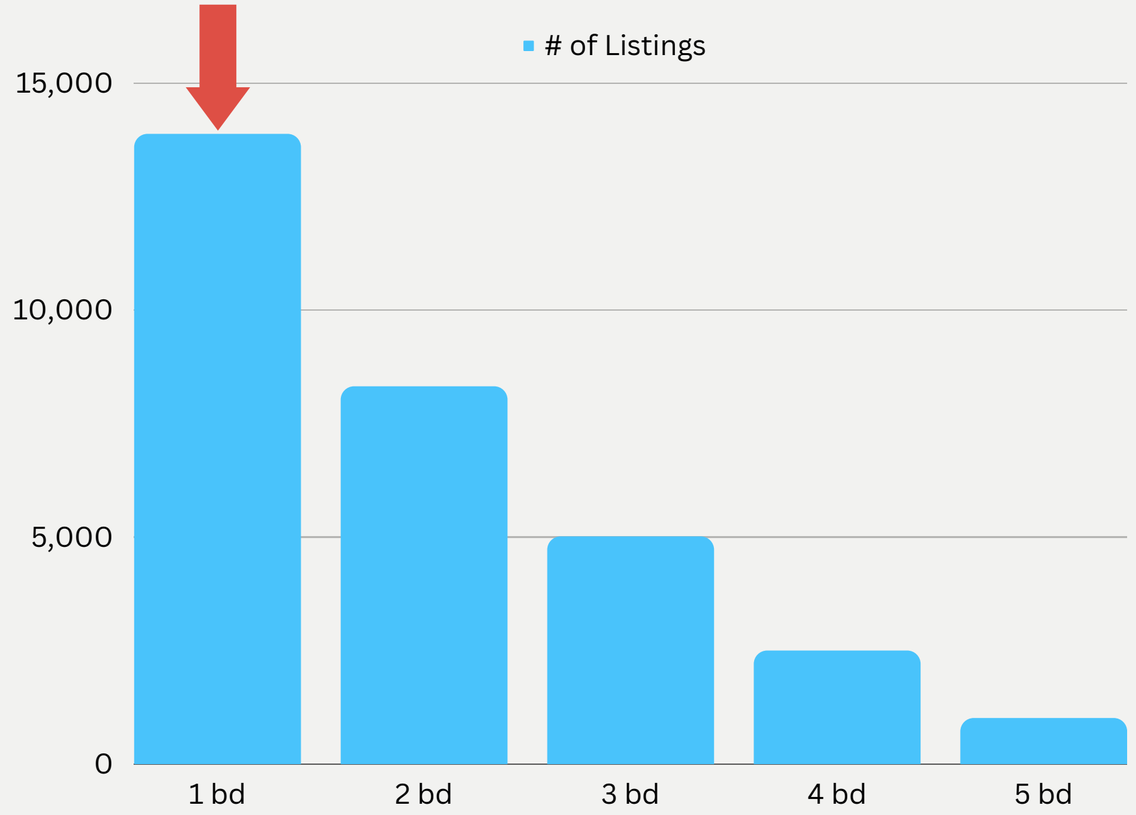
LA LISTINGS DATA: 44,000+ LISTINGS & 75 FEATURES

REVIEWS DATA: 1.5 MILLION REVIEWS

SOURCED FROM “INSIDEAIRBNB.COM”

DATA PROCESSING: LISTINGS

- ANALYSIS FOCUS: 1 BEDROOM LISTINGS
- DROPPED MONTHLY LISTINGS
- CONVERT PRICE TO DISCRETE TARGET VARIABLE: PRICE RANGE CLASSES



PRICE CLASS	PRICE RANGE
1	<$109
2	$109 - $149
3	$149 - $199
4	>$199

DATA PROCESSING - REVIEWS

listing_id	id	date	reviewer_id	reviewer_name	comments
109	74506539	2016-05-15	22509885	Jenn	Me and two friends stayed for four and a half ...
109	449036	2011-08-15	927861	Edwin	The host canceled my reservation the day befor...

1

GROUP REVIEWS ON ‘LISTING ID’

2

DROP LISTINGS WITH NO REVIEWS

3

MERGE ‘LISTINGS’ & ‘REVIEWS’ DATAFRAMES

1.5 million data points to 5,500!

DATA PROCESSING - NLP

- **DROP IRRELEVANT CHARACTERS**
  EX. ‘#’, ‘non-english characters’, ‘123’
- **PART OF SPEEAH TAGGING**
  (POS TAGGING)
- ***WORDNET LEMMATIZER***
- **VECTORIZE:**
  COUNT & TFIDF

EXAMPLE REVIEW:

“This is such a cute studio in a terrific LA location!”

REVIEW PROCESSED:

[‘cute’, ‘studio’, ‘terrific’, ‘la’, ‘location’]

EVALUATION METRIC

ACCURACY

HOW ACCURATELY MODEL PREDICTS PRICE CLASS

RECALL - FALSE NEGATIVE

PRICE INCORRECTLY CLASSIFIED

PRECISION - FALSE POSITIVE

MODELING:
3 GROUPS

1

FEATURES

REVIEWS ONLY (TEXT)

1. **Logistic Regression (Baseline Model)**
2. **Random Forest Classifier**
3. **K-Nearest Neighbors**
4. **Naive Bayes**
5. **Neural Network**

2

FEATURES

REVIEWS, RATINGS, NEIGHBORHOOD

1. **Logistic Regression**
2. **Random Forest Classifier**

3

FEATURES

REVIEWS, ALL RATINGS, NEIGHBORHOOD, PROPERTY TYPE & BATHROOMS

1. **Logistic Regression**
2. **Random Forest Classifier**

FINAL MODELS

LOGISTIC REGRESSION
UNTUNED

TRAIN
83%

TEST
55%

RANDOM FOREST CLASSIFIER
TUNED

TRAIN
84%

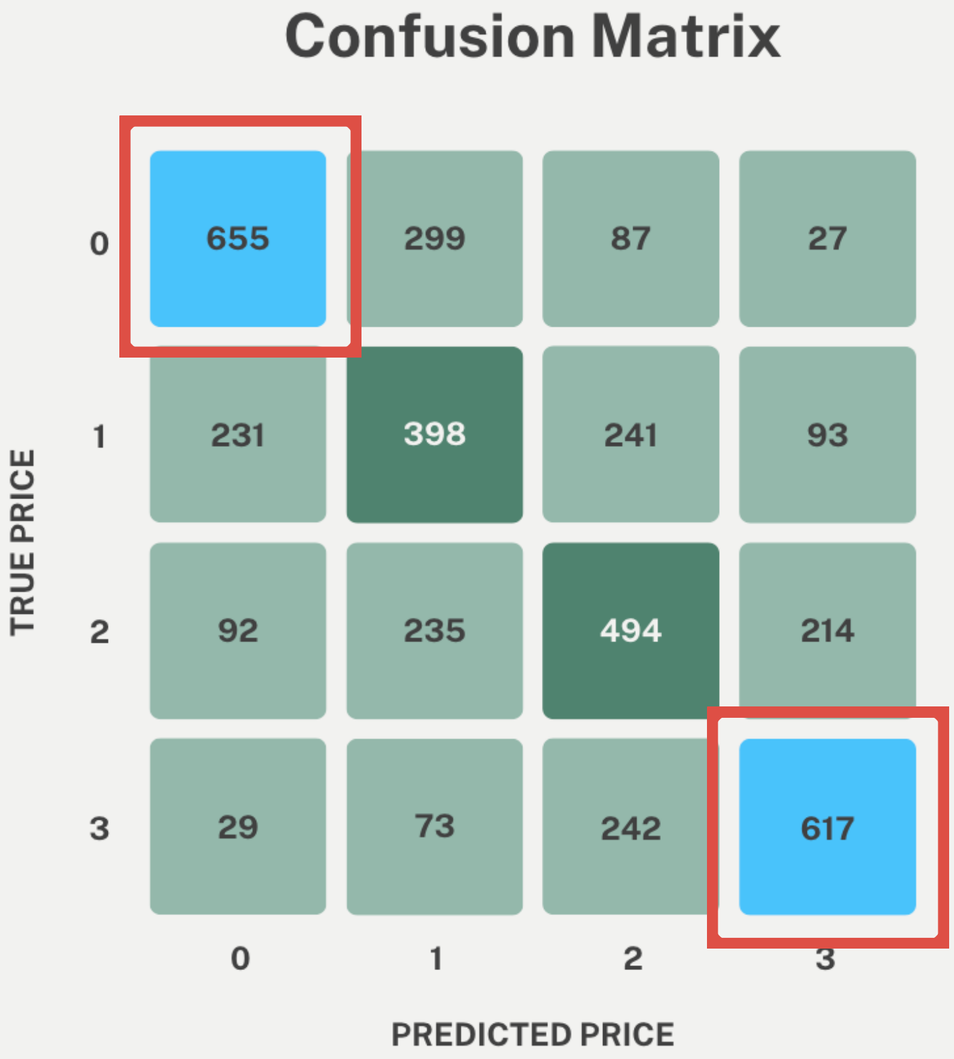
TEST
53%

EVALUATION

LOW OVERALL ACCURACY:
55%

PRICE CLASS 1 & 4:
ACCRUACY = ~70%

FURTHER TUNING COULD LEAD TO
A DECENT PREDICTOR



FEATURE IMPORTANCES (reviews)

RECOMMENDATIONS

- UNRELIABLE MODEL
- USE MODEL FOR SPECIFIC CLASSES
- POTENTIAL FOR GOOD PREDICTOR

NEXT STEPS

PARALLEL ANALYSIS:
REGRESSION

ADVANCED DATA PROCESSING:
‘BERT’ & ‘VADER’

THANK YOU

 jfbr1283

 jfwholehealth@gmail.com