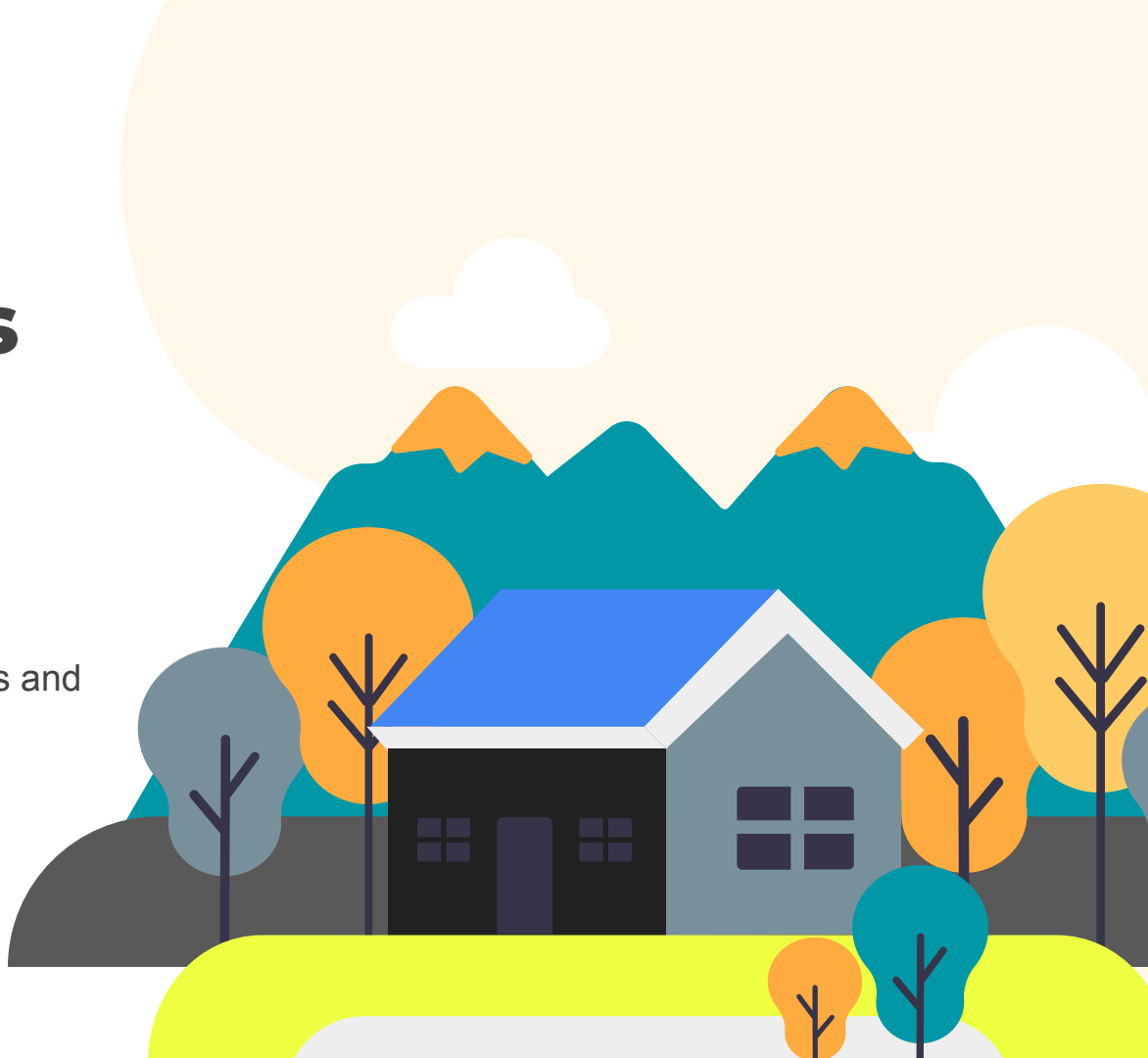# L^3: Listings Learning Labyrinth

Denver AirBnB Landscape Analysis and Predictions

EXECUTIVE SUMMARY

ANALYSIS OF AIRBNB DENVER PROPERTIES → **SEGMENTATION ANALYSIS** → **PREDICTIONS** → **RESULTS**

What patterns can machine learning identify within the inventory?

5 segments were identified within the inventory, with specific features in common

Can we predict overall rating, availability at 30 days, price and clusters?

Successfully predicted overall ratings, clusters/segments
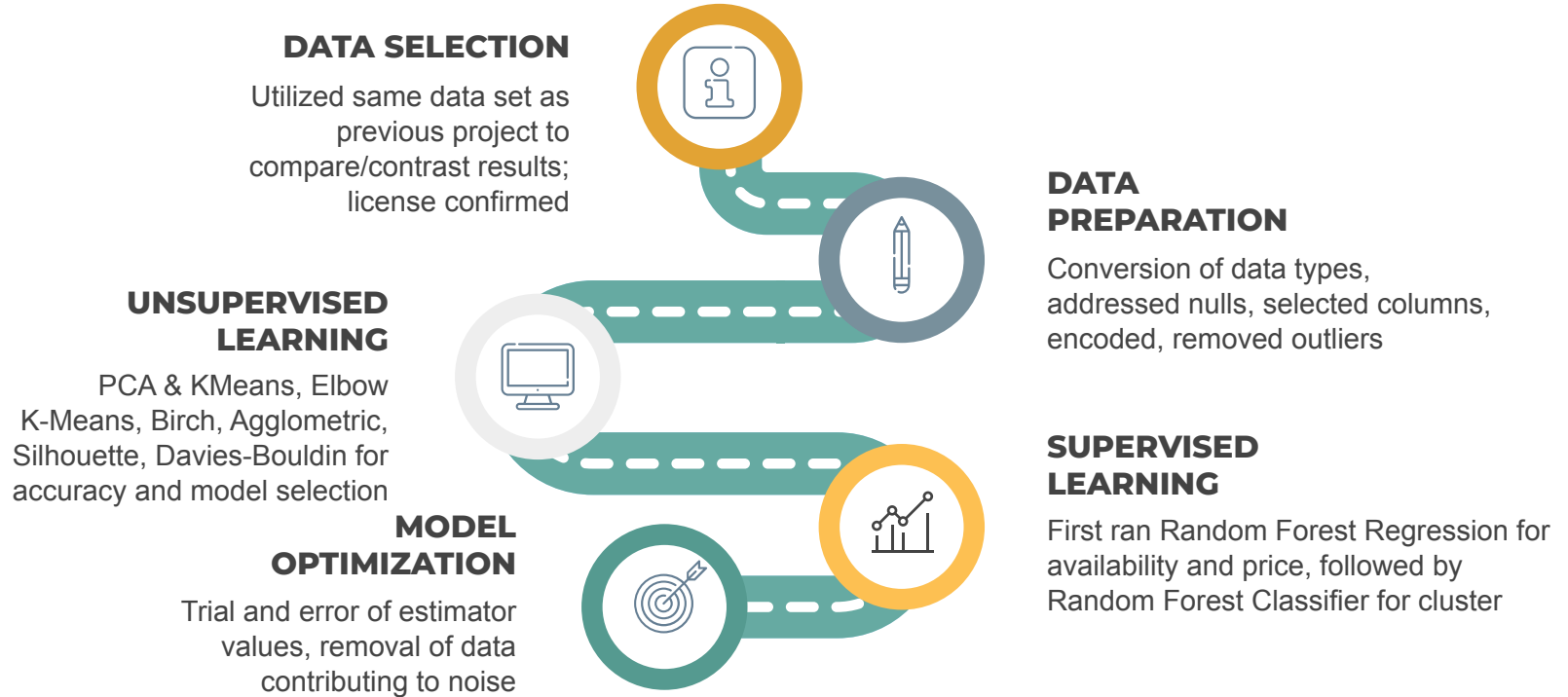
Identified feature importance for overall ratings

Models did not successfully predict price, 30 day availability

Data Source: *Inside Airbnb*
Inside Airbnb is a mission driven project that provides data and advocacy about Airbnb's impact on residential communities
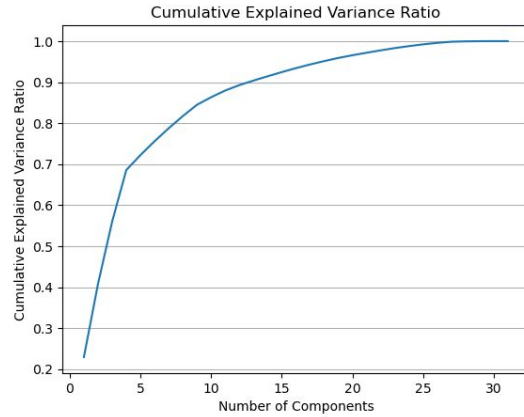*Listings and reviews through September 23, 2024*

DATA SELECTION
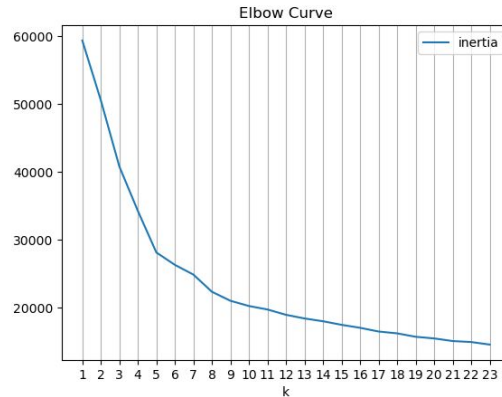
Utilized same data set as previous project to compare/contrast results; license confirmed

DATA PREPARATION

Conversion of data types, addressed nulls, selected columns, encoded, removed outliers

UNSUPERVISED LEARNING

PCA & KMeans, Elbow K-Means, Birch, Agglometric, Silhouette, Davies-Bouldin for accuracy and model selection

SUPERVISED LEARNING

First ran Random Forest Regression for availability and price, followed by Random Forest Classifier for cluster

MODEL OPTIMIZATION

Trial and error of estimator values, removal of data contributing to noise
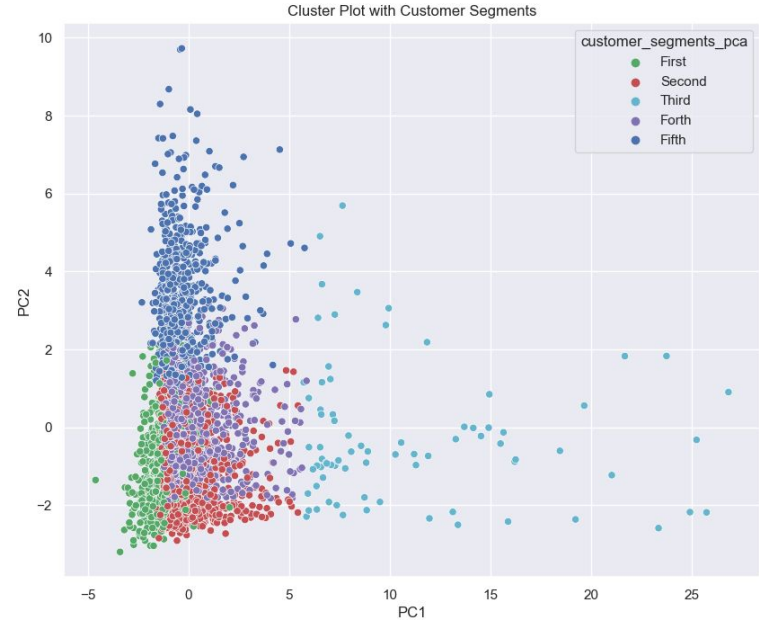
- PCA and K-means were used to create a five cluster model
- Silhouette Score: .26
- Davies-Bouldin Index: 1.36

Explained Variance Ratio

Clusters

Elbow Curve

SEGMENTATION ANALYSIS: 5 CLUSTERS IDENTIFIED

#0

N = 614

No. of **Reviews**

Property type: guest suite & guest house

#1

N=1202

Had the **most availability** 30 days out

Property type: entire condo

#2

N = 78

Lowest review **ratings** and # of reviews
Property type: rental units

#3

N= 983

Had the **least availability** 30 days out

Property type: condo and other types of properties

#4

N = 661

No. of **baths**, **bedrooms** and highest **price**

Property type: entire home or townhouse

SUPERVISED LEARNING MODELS

WHICH GROUP?

30d AVAILABILITY

OVERALL RATING

PRICE

Random Forest Classifier

HIGH PERFORMING Accuracy Score= .895

Random Forest Regressors

LOW PERFORMING MAE= 7.3   RMSE= 9
MSE= 81.2   R2= .02

HIGH PERFORMING MAE= .06   RMSE= .13
MSE= .018   R2= .80

LOW PERFORMING MAE= 42.7   RMSE= 56.8
MSE= 3228.4   R2= .47

POSSIBLE APPLICATIONS
- With further analysis could be leveraged for recommendation engine (renter or owner)
- Automate future cluster assignments
- Support marketing campaigns

POSSIBLE APPLICATIONS
- Offer guidance to hosts on how to improve overall rating
- Leveraged for recommendation engine (renter)

```
Classification Report
              precision    recall  f1-score   support

           0       0.90      0.81      0.85       178
           1       0.89      0.91      0.90       277
           2       0.92      0.71      0.80        17
           3       0.86      0.91      0.88       237
           4       0.94      0.96      0.95       183

    accuracy                           0.90       892
   macro avg       0.90      0.86      0.88       892
weighted avg       0.90      0.90      0.90       892
```
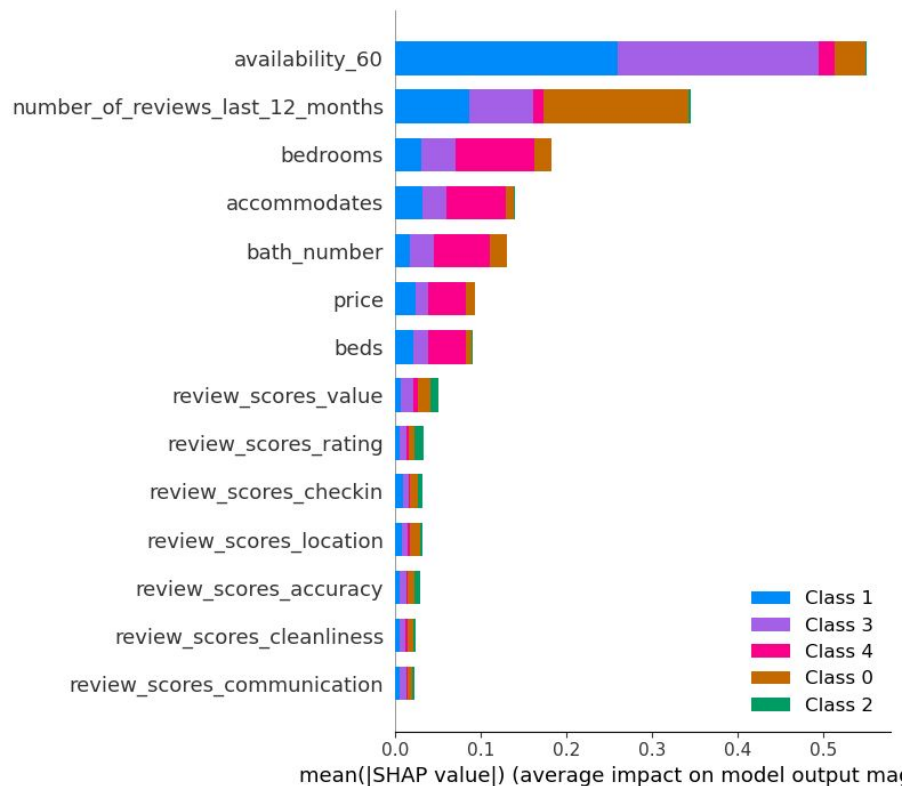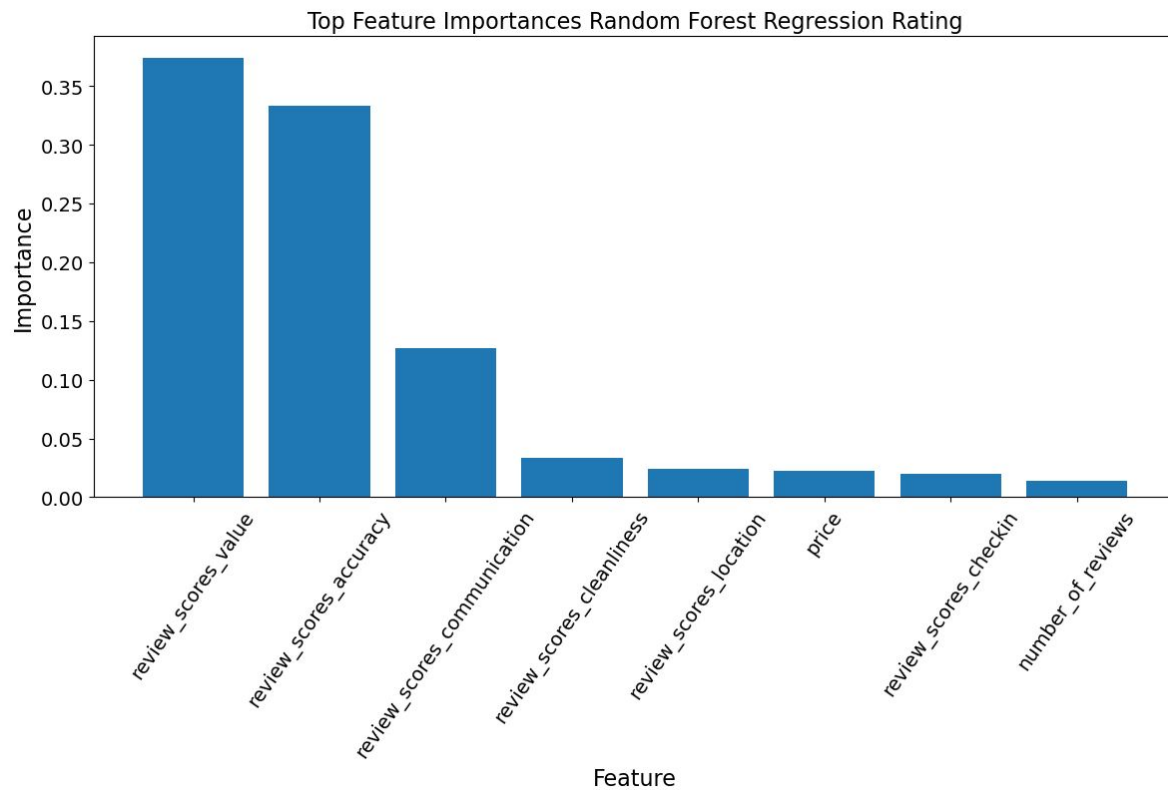
DID WE OVERFIT?

```
Accuracy Score Training Data: 1.0
Accuracy Score Testing Data: 0.8957399103139013
```


mean(|SHAP value|) (average impact on model output magnitude)

Legend: Class 1, Class 3, Class 4, Class 0, Class 2

Features (top to bottom): availability_60, number_of_reviews_last_12_months, bedrooms, accommodates, bath_number, price, beds, review_scores_value, review_scores_rating, review_scores_checkin, review_scores_location, review_scores_accuracy, review_scores_cleanliness, review_scores_communication

Top Feature Importances Random Forest Regression Rating

THANKS

Does anyone have any questions?