# INTERPRETABLE ML- KAGGLE COMPETITION

GROUP B

NITISH ACHARYA , JULIA BLUME, ELIAS BRUMMUND, AYREEN JAPUTRI, VAHE SHELUNTS

# WHAT DATA SCIENTISTS LOOK LIKE

NITISH

Expert for
deriving addresses

ELIAS

Expert for
SHAP Values

AYREEN

Expert for
Decision Trees

JULIA

Expert for
Partial Dependence
Plots

VAHE

Expert for
Feature Engineering

# FROM RAW DATA TO KNOWLEDGE


EXPLORATORY DATA ANALYSIS

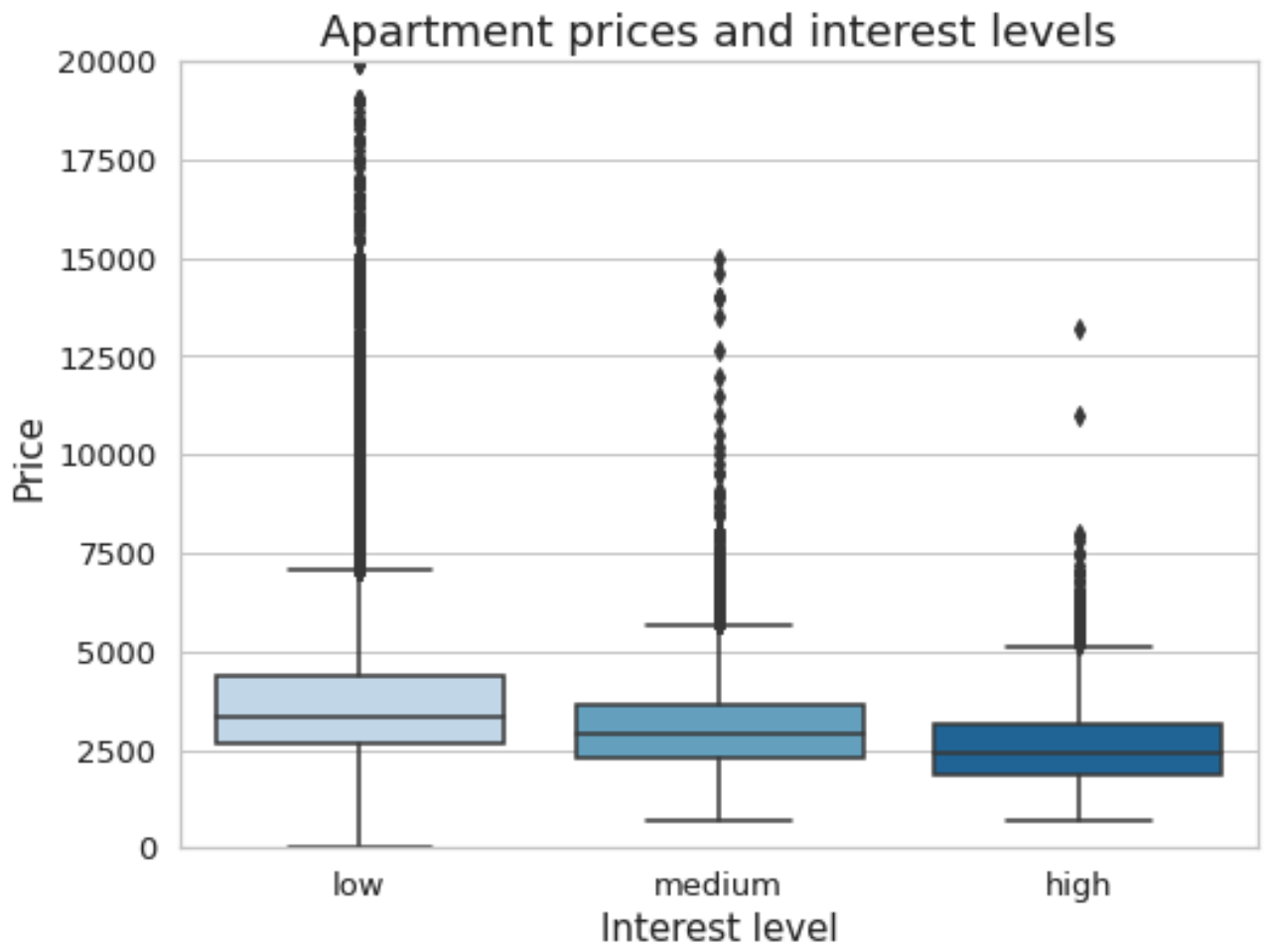
PREPARATION

MODELLING

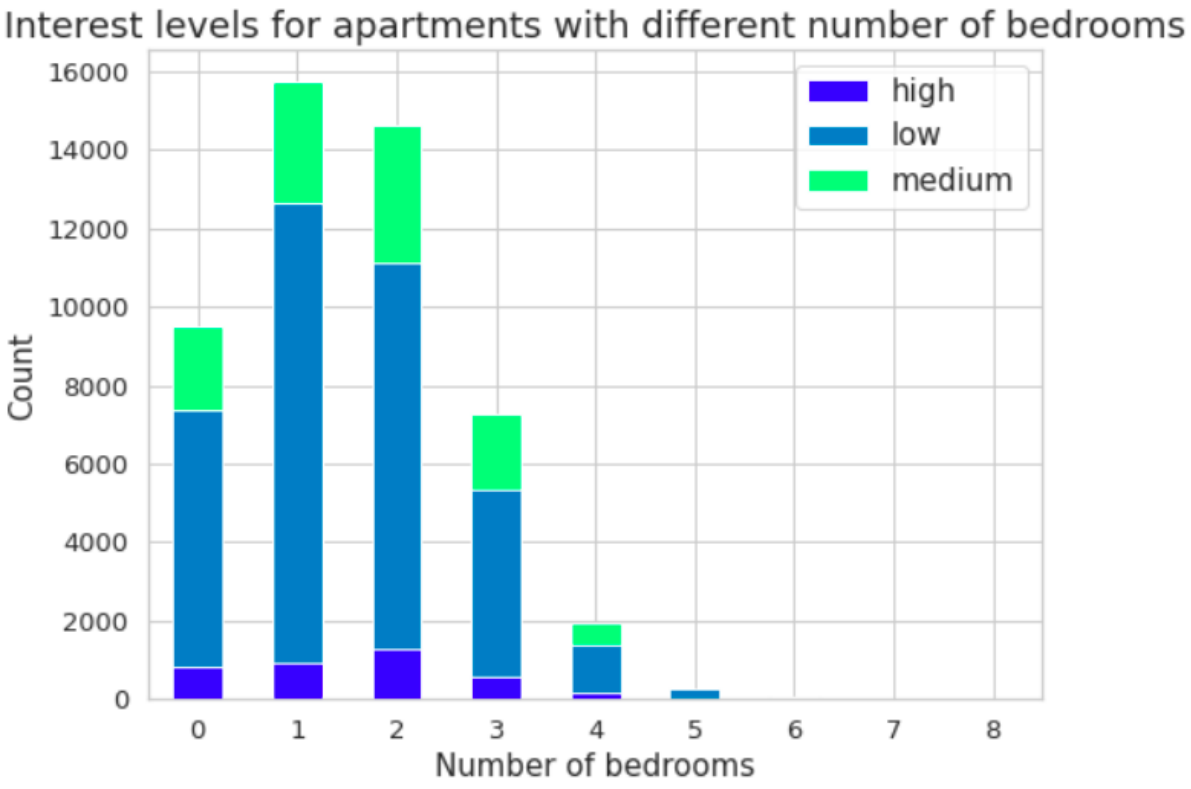
EXPLANATION

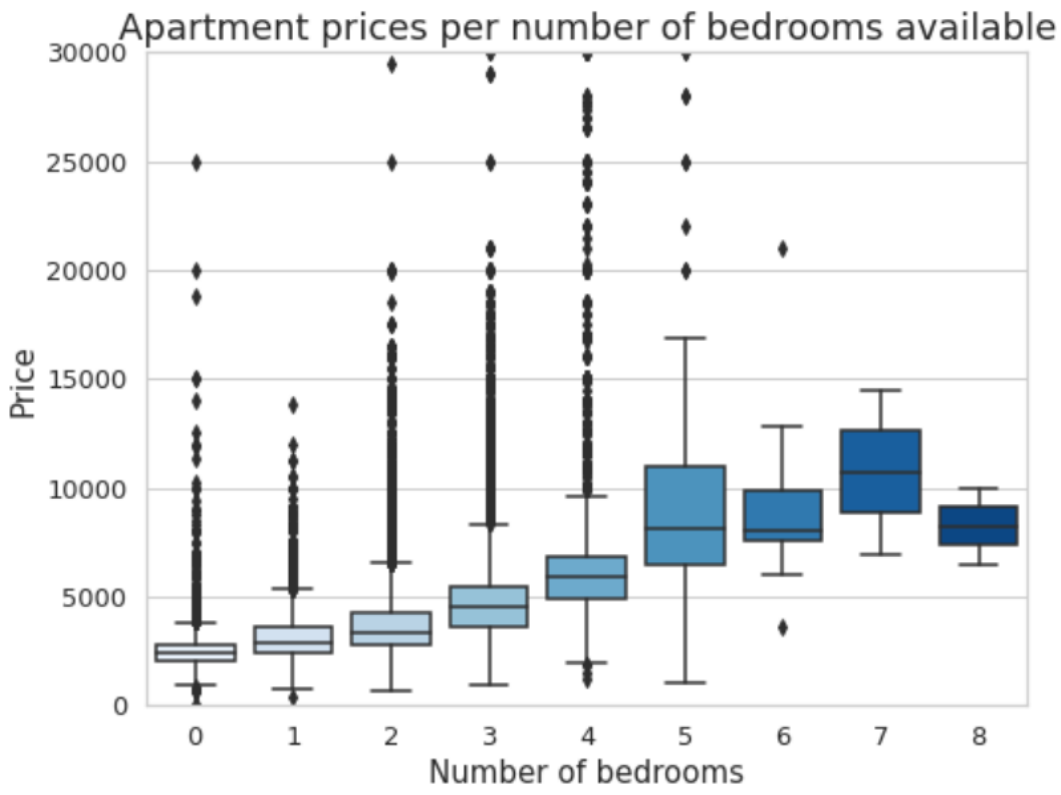
# WHAT WAS SPECIAL ABOUT THE DATA SET - FINDINGS FROM EDA

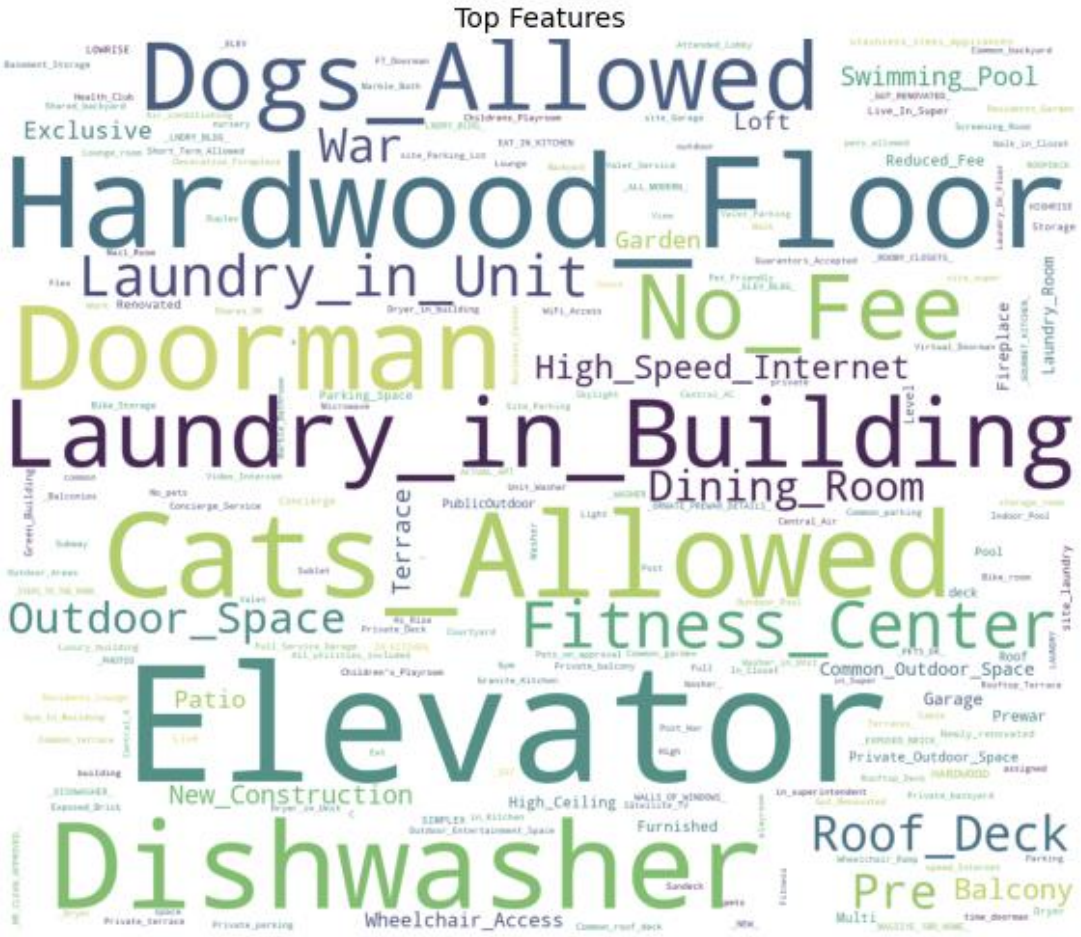
Increasing interest level as the apartment price decreases



# WHAT WAS SPECIAL ABOUT THE DATA SET






Word cloud for most used words in column “features“

# FEATURE ENGINEERING ON THE COLUMN “FEATURES”

# FEATURE ENGINEERING ON THE COLUMN "FEATURES"

```
special_char = ['!', '"', '$', '&', '(', ')', '*', '+', '-', '.', '/', ':', ';', '<', '>', '@', '^', '~', '\xa0', '®', '•']
for x in range(len(special_char)):
    match = [e for e in feature_total if special_char[x] in e]
    print('First 5 matches for ', special_char[x], ': ', match[:5], '\n')
```

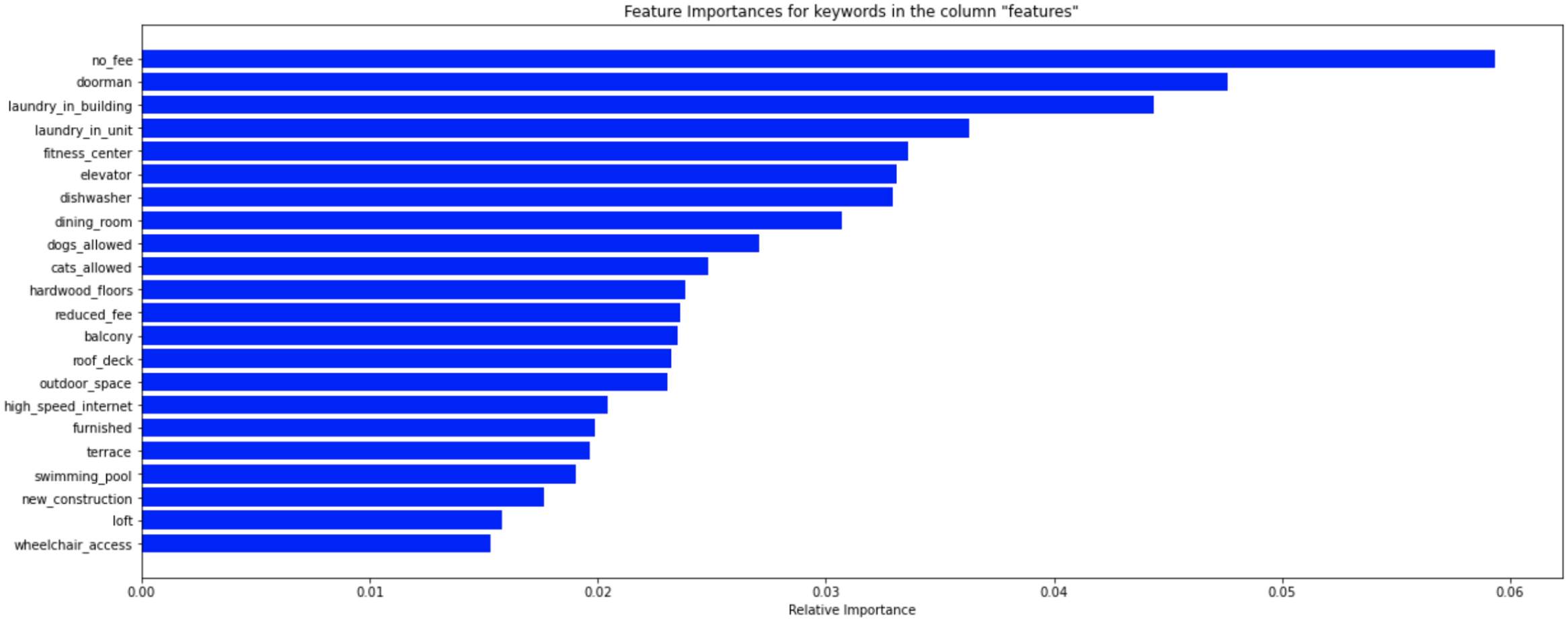
```
First 5 matches for  ! :  ['** HOLY NO FEE DEAL BATMAN! * OVERSIZED 2BR HOME * SPARKLING CLEAN & BRITE * HEART OF GREENPOINT * NEAR THE PARK & TRAIN
S **', '!!!!LOW FEE!!!!', '** CHELSEA BABY! * MASSIVE 2BR SUPER SHARE * ALL MODERN & NEW * ELEV /LNDRY BLDG **', 'Garage Parking!', '** COURT SQUARE
GEM! * SPRAWLING SUNDRENCHED 2BR HOME * CUSTOM FINISHES * DISHWASHER * FIREPLACES * EAT-IN KITCHEN * BAY WINDOWS **']
```

Clean words from special characters to make them comparable

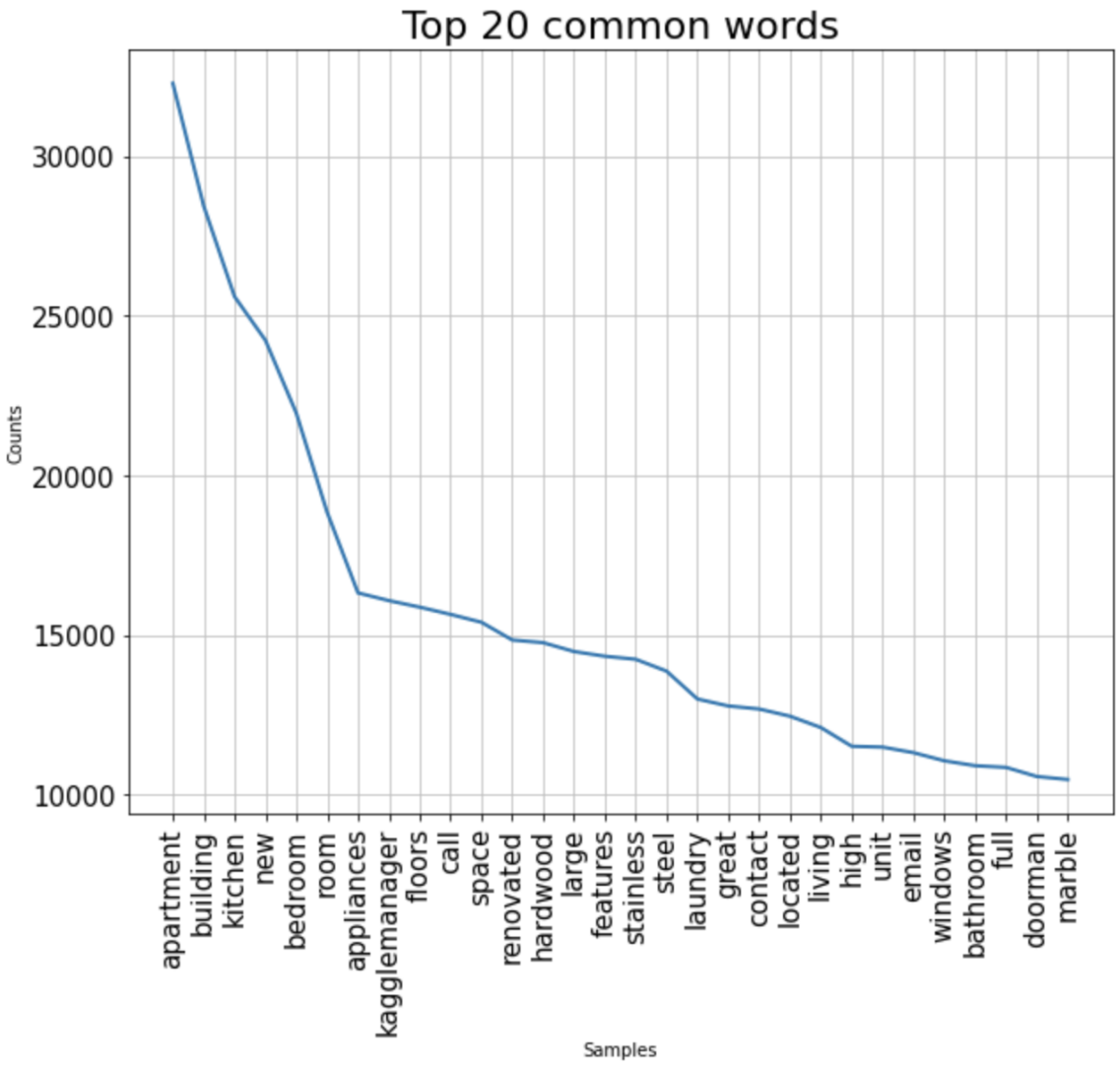
Encode keywords into a data frame, add interest level column and fit a random forest

Compute feature importance to find TOP 10 keywords

# FEATURE ENGINEERING ON THE COLUMN “FEATURES”



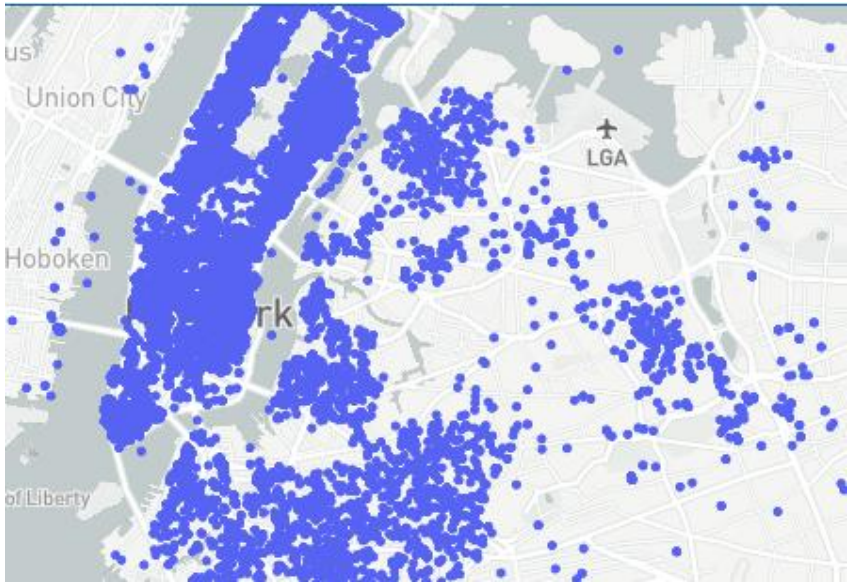
# MAKING USE OF THE DESCRIPTION COLUMN



# FROM LONGITUDE/ LATITUDE TO AN ADDRESS

? To see if the price and interest level are dependent upon a particular area

1. geopy lib
2. create service provider locator
3. timeout =10 sec
4. reverse geocoding

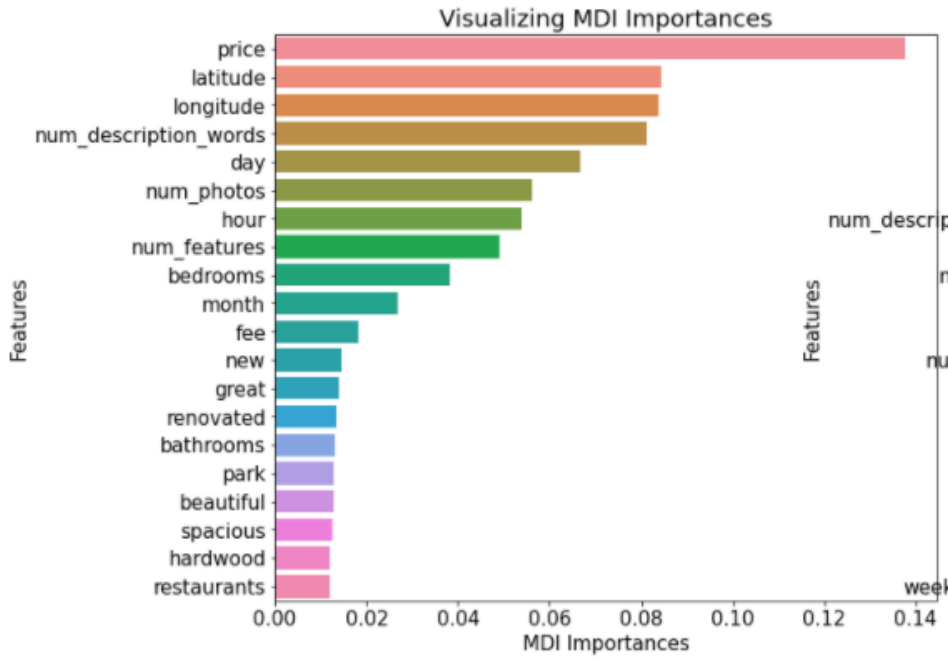


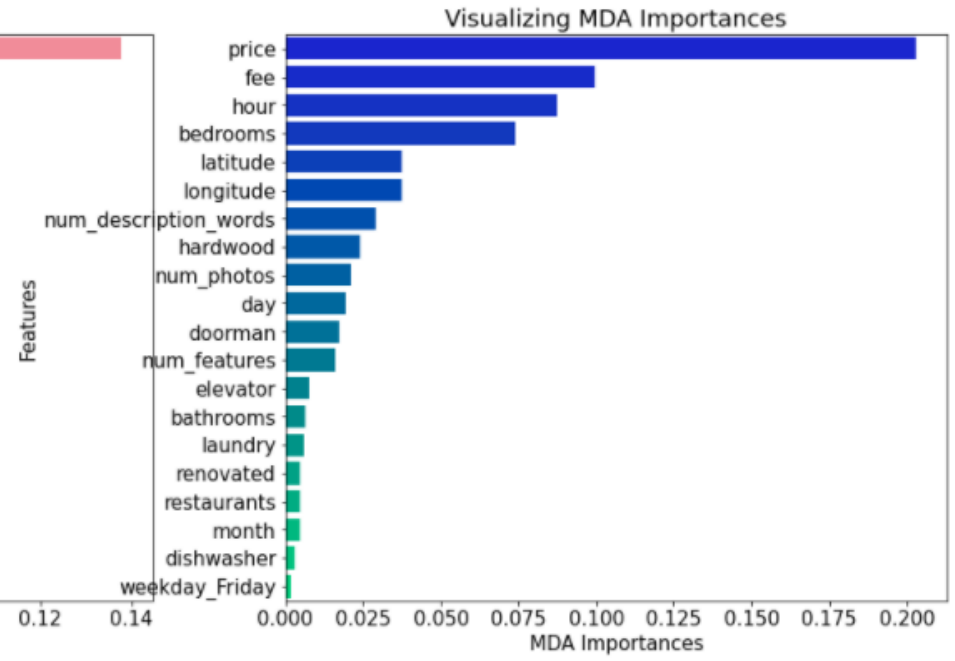

| interest_level | price | listing_id | Streetnumber | Streetname | neighbourhood | Community |
|---|---|---|---|---|---|---|
| high | 3195 | 6811957 | 752 | Broadway | NoHo | NoHo Historic District |
| medium | 2000 | 6811965 | 230 | East 54th Street | Midtown East | Manhattan Community Board 6 |
| high | 5850 | 6811966 | 135 | East 22nd Street | Gramercy | Manhattan Community Board 6 |
| medium | 2745 | 6811973 | 269 | West 94th Street | Upper West Side | Manhattan Community Board 7 |

# FINAL DATAFRAME FOR MODELLING


| # | Column |
|---|---|
| 0 | bathrooms |
| 1 | bedrooms |
| 2 | latitude |
| 3 | longitude |
| 4 | price |
| 5 | interest_level |
| 6 | num_photos |
| 7 | num_features |
| 8 | num_description_words |
| 9 | hardwood |
| 10 | doorman |
| 11 | fee |
| 12 | cats |
| 13 | laundry |
| 14 | war |
| 15 | fitness |
| 16 | elevator |
| 17 | dishwasher |
| 18 | dogs |
| 19 | The Bronx |
| 20 | Lenox Hill |
| 21 | Queens |
| 22 | Stuy Town |
| 23 | Bay Ridge |
| 24 | Steinway |
| 25 | Sunset Park |
| 26 | Upper West Side |
| 27 | Gramercy |
| 28 | Park Slope |
| 29 | new |
| 30 | renovated |
| 31 | highlarge |
| 32 | great |
| 33 | restaurants |
| 34 | park |
| 35 | spacious |
| 36 | beautiful |
| 37 | access |
| 38 | center |
| 39 | year |
| 40 | month |
| 41 | day |
| 42 | hour |
| 43 | weekday_Friday |
| 44 | weekday_Monday |
| 45 | weekday_Saturday |
| 46 | weekday_Sunday |
| 47 | weekday_Thursday |
| 48 | weekday_Tuesday |
| 49 | weekday_Wednesday |

# RANDOM FOREST



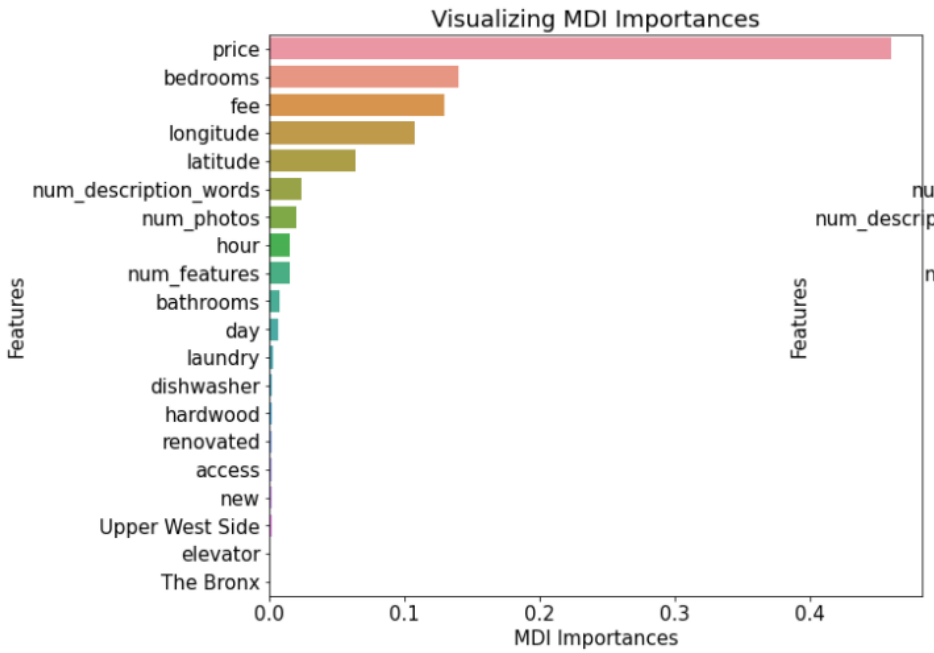

## PRUNING

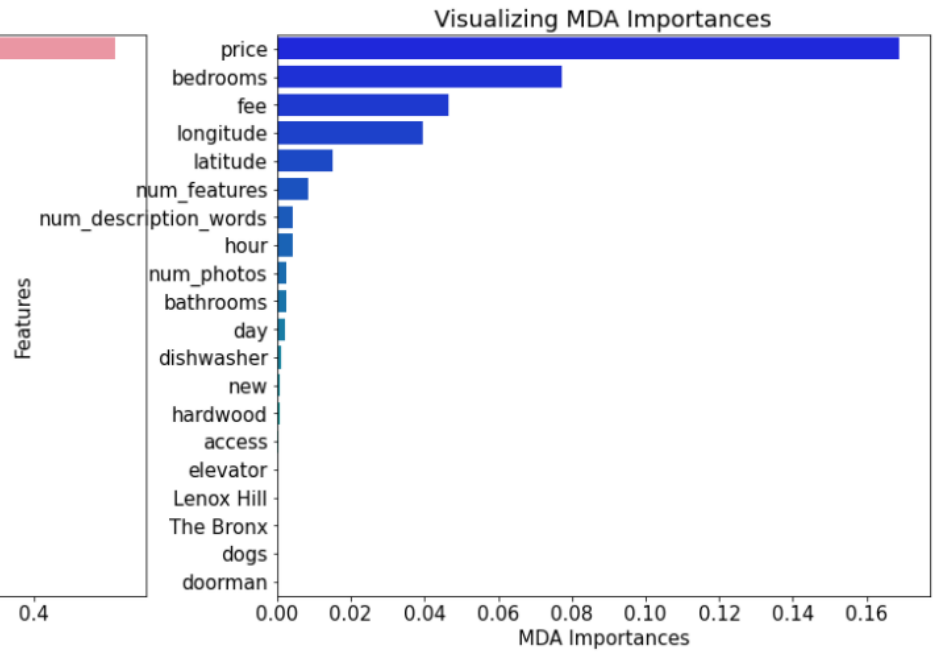
- Scikit-Learn RandomizedSearchCV method

## HYPERPARAMETERS

| | |
|---|---|
| n_estimators: | **144** |
| min_samples_split: | **10** |
| min_samples_leaf: | **1** |
| max_features: | **sqrt** |
| max_depth: | **60** |
| bootstrap: | **False** |

# DECISION TREE




## PRUNING

- Scikit-Learn RandomizedSearchCV method

## HYPERPARAMETERS

| | |
|---|---|
| max_depth: | **25** |
| max_features: | **36** |
| min_samples_leaf: | **41** |

# RANDOM FOREST VS DECISION TREE

**SCORES**

| | | | |
|---|---|---|---|
| CROSS ENTROPY: | 0.702 | ↘ | 0.679 |
| ACCURACY: | 0.691 | ↘ | 0.688 |
| RECALL: | 0.691 | ↘ | 0.688 |
| ROC AUC: | 0.743 | ↗ | 0.750 |

**SCORES**

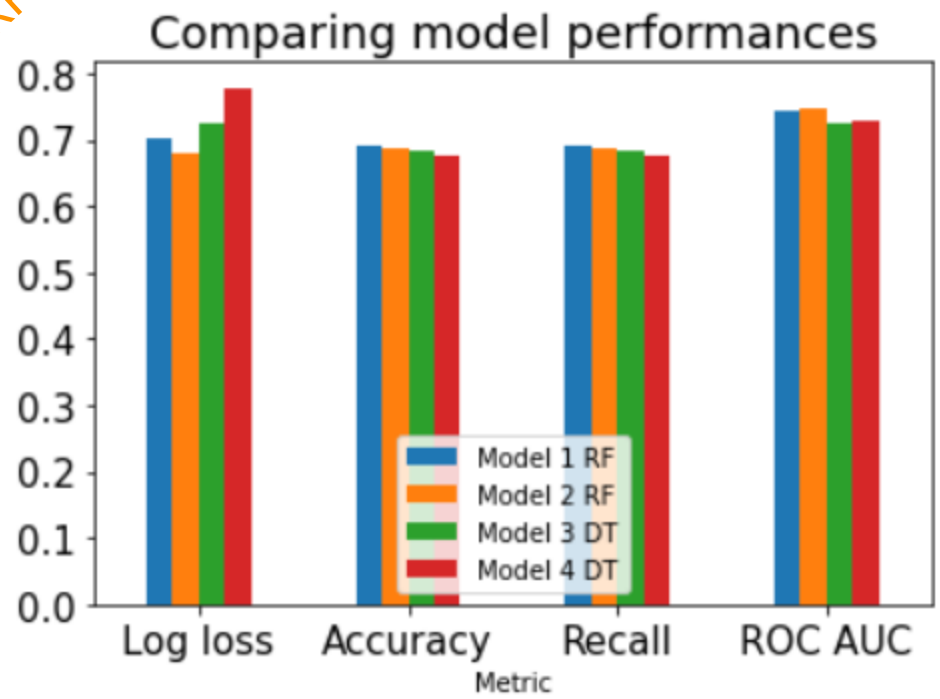
| | | | |
|---|---|---|---|
| CROSS ENTROPY: | 0.724 | ↗ | 0.778 |
| ACCURACY: | 0.682 | ↘ | 0.677 |
| RECALL: | 0.682 | ↘ | 0.677 |
| ROC AUC: | 0.724 | ↗ | 0.727 |

1: plain RF

2: pruned RF

3: plain DT

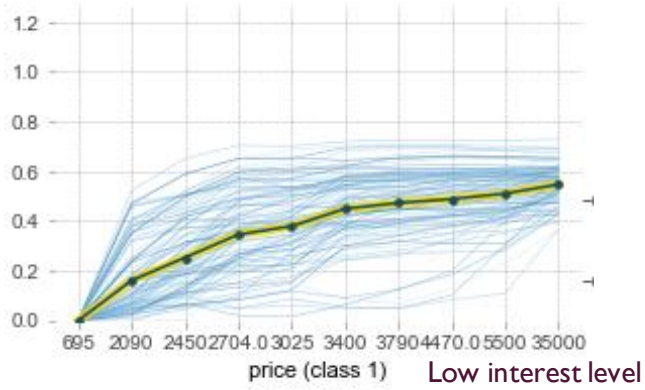
4: pruned DT



# PARTIAL DEPENDENCE PLOTS

## 1 FEATURE – 2 MODELS

Feature: price
Model: Random Forest

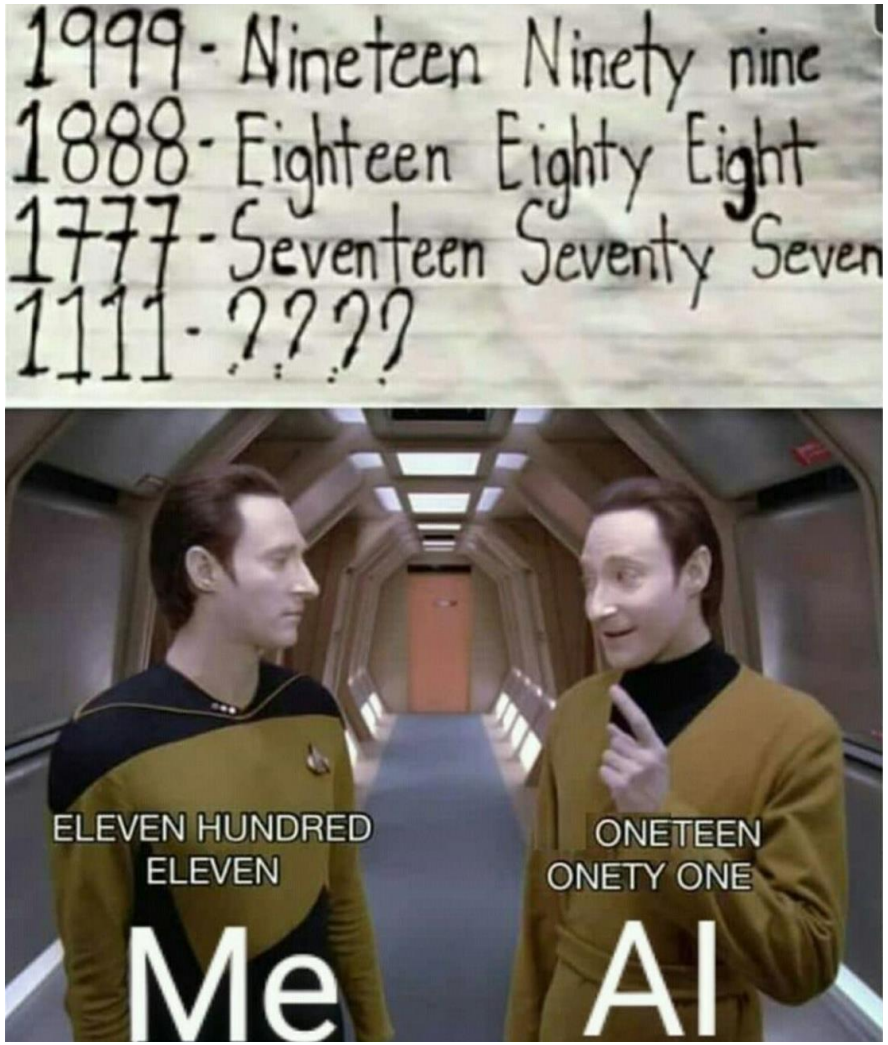



Feature: price
Model: Decision Tree

# EXPLAINING A MODEL

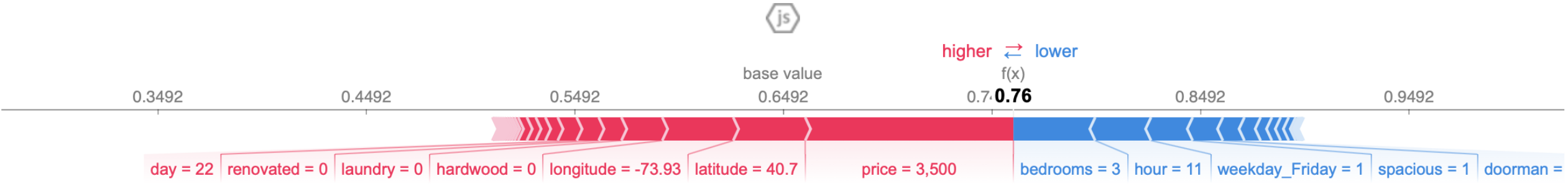
"CAN I TRUST

THE PREDICTION?"

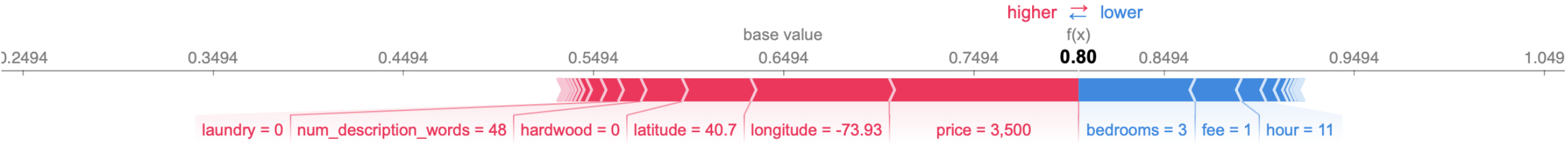


# LOCAL EXPLANATION WITH SHAP PLOT

## Random Forest




## Decision Tree




# PROJECT TAKE AWAYS

"**Impressed** by how much we found out in only 4 weeks"

"is that enough **text**? I mean everyone can read the code?"

"some data (columns) **looks simple but** using it requires a ton of work"

"complicated to find a suited platform for **group coding** "

"can you explain **SHAP** one more time?"

"It's **NOT just 3** lines of code"

„**validation** or **train** set?"