Capstone AirBnB ML Solution Presentation

Team 8 - Group Introductions

Adetoso Afonja

Duong Dinh

Ella Elazkany

Faraz Khadivpour

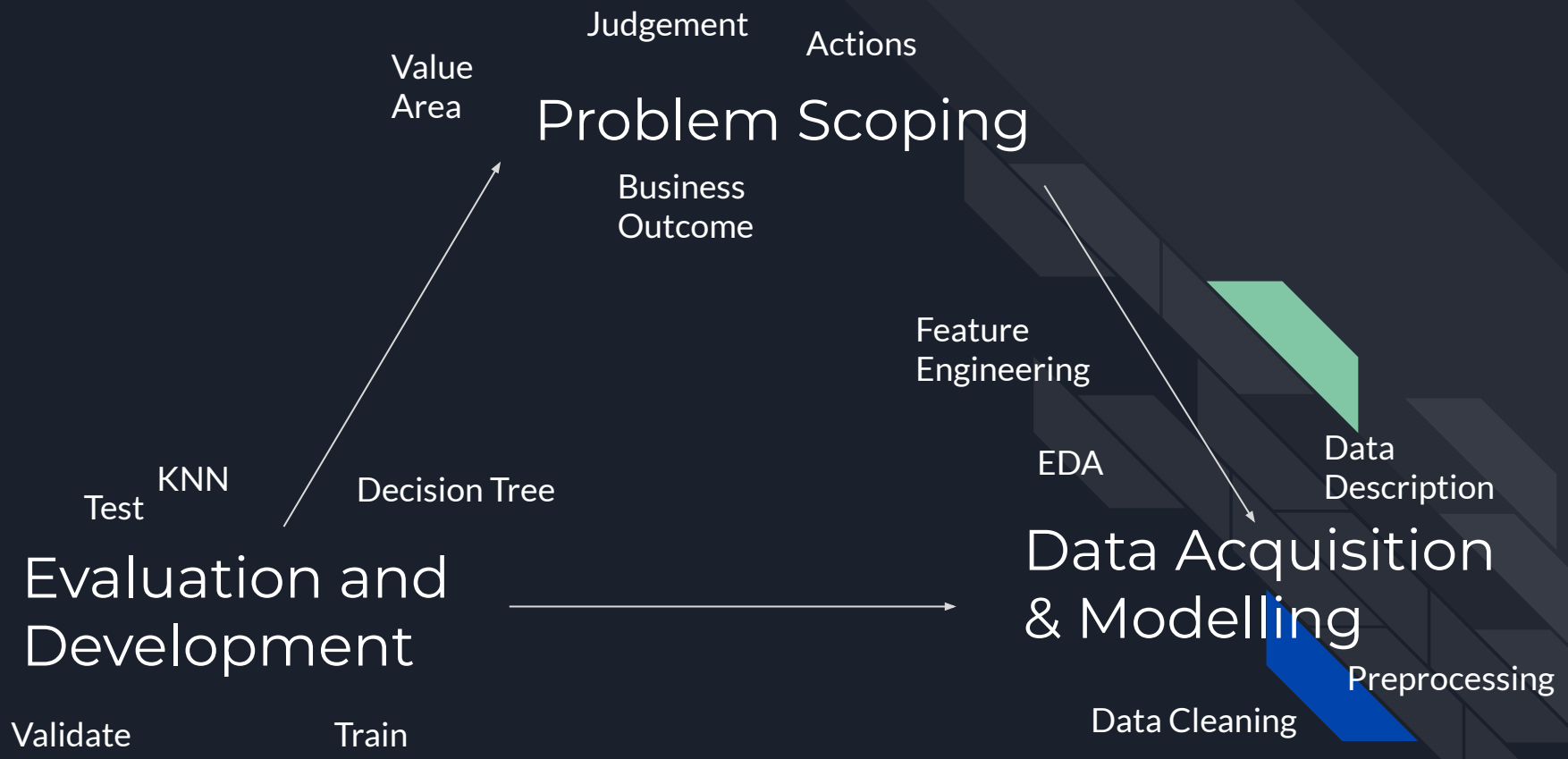
Business Overview - AirBnB

- AirBnB is a vacation rental marketplace company.
- It maintains and hosts a marketplace accessible to consumers on its website or app.
- Through its service users can arrange lodging, primarily home stays and tourism experience or list of their properties for rentals.

Business Overview - AirBnB

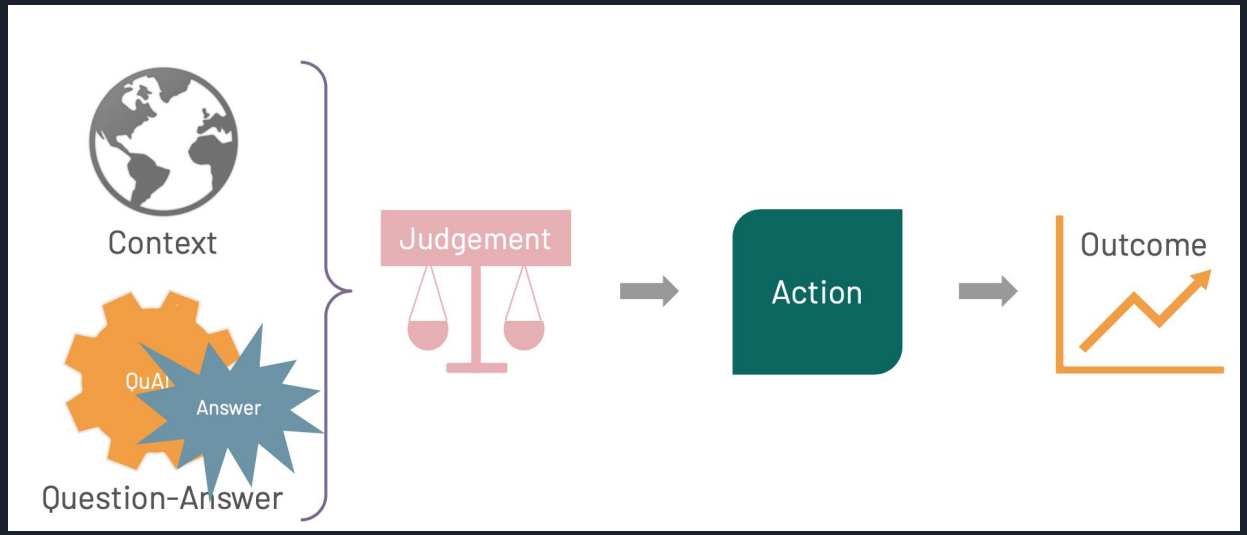
- There are 2.9 million host on Airbnb worldwide in 2020
- 14,000 new hosts are joining the platform each month in 2020
- There are over 7 million listings on Airbnb worldwide in 2020
- There are 100,000 cities with active Airbnb listings in 2020
- There are 220 countries and regions with active Airbnb listings in 2020

Source:
https://www.stratosjets.com/blog/airbnb-statistics/#:~:text=According%20to%20Airbnb%20data%20there,active%20Airbnb%20listings%20in%202020.


Judgement
Actions
Value
Area
Problem Scoping
Business
Outcome
Feature
Engineering
EDA
Data
Description
Data Acquisition
& Modelling
Preprocessing
Data Cleaning
Test
KNN
Decision Tree
Evaluation and
Development
Validate
Train

Problem Scoping

Problem Scoping Process



Problem Scoping Backwards Process - Outcomes

Initial Brainstorming Outcomes

- Improve hosting Air B&B experience
- Improving customer Air B&B experience
- Improve customer engagement through review
- Optimize room prices for host
- Optimize room prices for customers
- Correlation and prediction of prices per area

Target Outcome

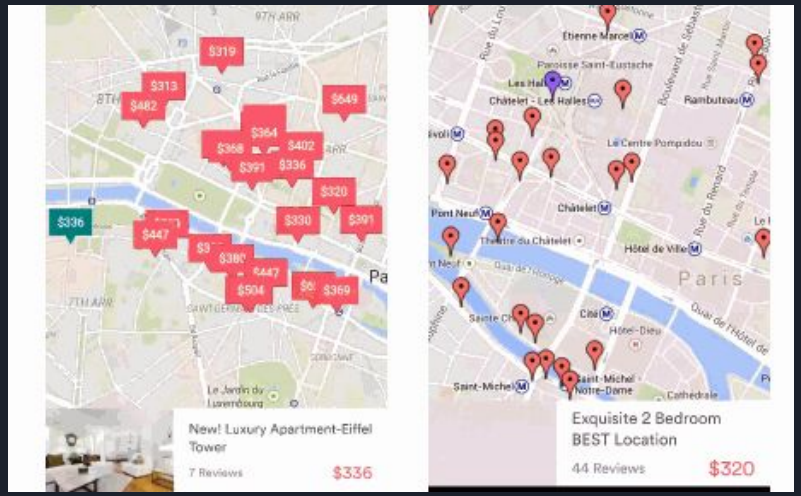
Determine the **optimal nightly rent price**.

Why? Because, based on the marketing data (features), if rent price is above market price, renters will select more affordable alternatives and hosts won't make money. On the other hand, if rent price is too low, hosts will miss out on potential profit.

Problem Scoping Backwards Process - Actions

Action

- Improving host revenue: Recommending to the host a range and ranking of prices based off of their surrounding market conditions
- When a guest queries a room, their search results should promote rooms based off of location and market prices with simular vacancies, while maximizing fair market prices.



Problem Scoping Backward Process - Judgment Analysis

- If our QuAM produced a wrong recommendation and the price was high this, would in theory reduce the chances of the room being rented out
- If our QuAM produced a low wrong recommendation and the price was low , this would affect revenue for AirBnB as a business and the host.
- Host fees generally starts at 3% but can go up to 16% depending on the criterias of accomodation, losses in revenue can be significant at this rate

Problem Backwards Process - Questions Answer Pair

Question Template

- What is the optimal price should _____ at this location?

Object Template

- The Guest Pay
- The Room Be

Problem Scoping Forward Process - Looking at the Data

host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365
John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21	6	365
Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.38	2	355
Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	NaN	1	365
LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05	4.64	1	194
Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19	0.10	1	0

Problem Backwards Process - Questions Answer Pair

Question Template

- What is the optimal price should _____ at this location?

Object Template

- The Guest Pay
- The Room Be

Problem Scoping Forward Process - Questions Answer Pair

Question Template

- What is the optimal price **RANGE** should _____ at this location?

Object Template

- ~~The Guest Pay~~
- The Room Be

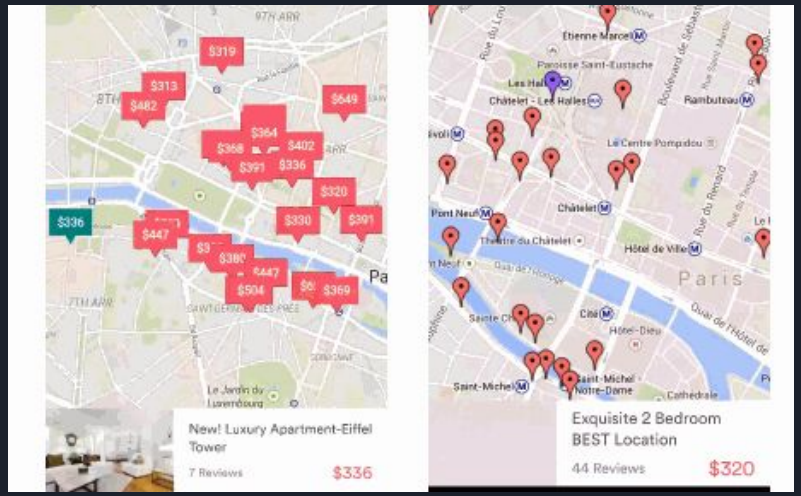
Problem Scoping Forward Process - Judgement

- If our QuAM produced a wrong rating and the price was high this, would in theory reduce the chances of the room being rented out
- If our QuAM produced a low rating recommendation and the price was low , this would affect revenue for AirBnB as a business and the host.

Problem Scoping Backwards Process - Actions

Action

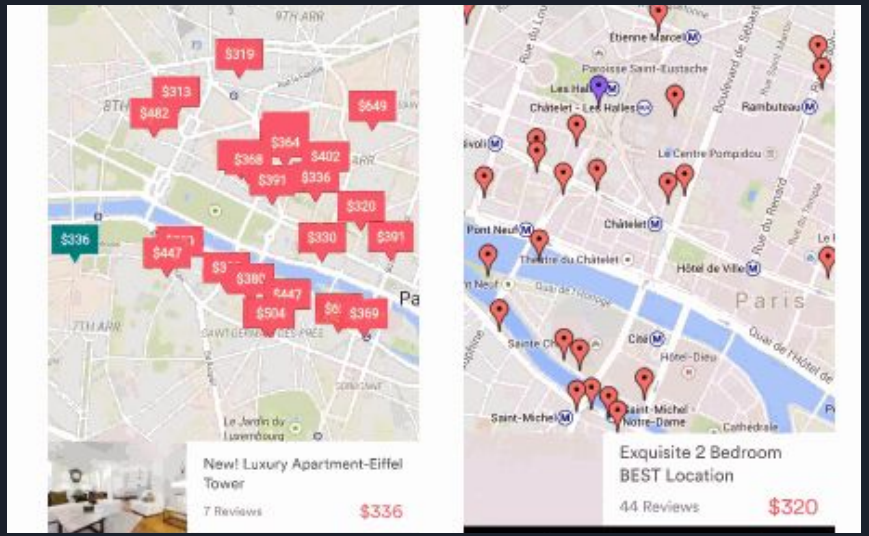
- Improving host revenue: Recommending to the host a range and ranking of prices based off of their surrounding market conditions
- When a guest queries a room, their search results should promote rooms based off of location and market prices with simular vacancies, while maximizing fair market prices.



Problem Scoping Backwards Process - Actions

Action

- Improving host renenue: Recommending to the host a range and ranking of prices based off of their surrounding market conditions



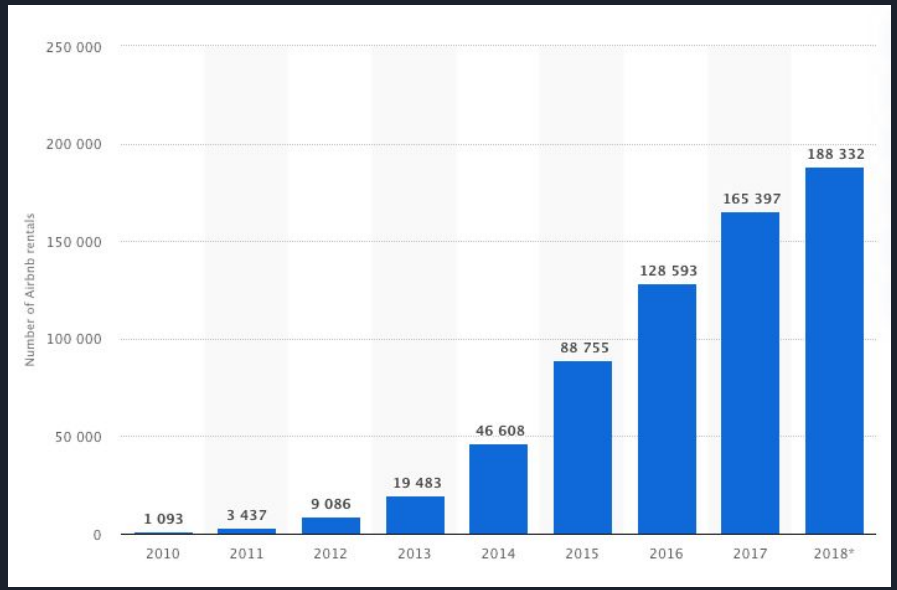
Value Areas

Projection

Project if the price of the host is in estimated market range using surrounding market data.

Performance Time & Cost

Number of Airbnb rental properties in New York in the United States from 2010 to 2018



For a company to be able to achieve what our QuAM does, it would require at team of people full-time to process this data. This would cost a company like AirBnB hundreds of thousands of dollars if not millions.

https://www.statista.com/statistics/915161/new-york-airbnb-rentals-united-states/#:~:text=The%20statistic%20depicts%20the%20number,Airbnb%20rentals%20in%20New%20York.

Business Outcome - Project Objective

- Determine the range of optimal nightly rental price based on the surrounding area
- Produce a rating or a range of prices based off of surrounding market conditions

Data Acquisition and Modelling

New York City Airbnb Open data

Open dataset downloaded from Kaggle
https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

Includes information about hosts, locations, availability, reviews, prices, …

All information are represented in columns

All columns are our feature variables except ‘price’ which is our target variable

Features in our dataset are either numerical or categorical

Exploring & cleaning Airbnb dataset

Exploring & cleaning

Remove duplicates & unnecessary features

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21	6	365
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.38	2	355
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	NaN	1	365
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05	4.64	1	194
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19	0.10	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
48890	36484665	Charming one bedroom - newly renovated rowhouse	8232441	Sabrina	Brooklyn	Bedford-Stuyvesant	40.67853	-73.94995	Private room	70	2	0	NaN	NaN	2	9
48891	36485057	Affordable room in Bushwick/East Williamsburg	6570630	Marisol	Brooklyn	Bushwick	40.70184	-73.93317	Private room	40	4	0	NaN	NaN	2	36
48892	36485431	Sunny Studio at Historical Neighborhood	23492952	Ilgar & Aysel	Manhattan	Harlem	40.81475	-73.94867	Entire home/apt	115	10	0	NaN	NaN	1	27
48893	36485609	43rd St. Time Square-cozy single bed	30985759	Taz	Manhattan	Hell's Kitchen	40.75751	-73.99112	Shared room	55	1	0	NaN	NaN	6	2
48894	36487245	Trendy duplex in the very heart of Hell's Kitchen	68119814	Christophe	Manhattan	Hell's Kitchen	40.76404	-73.98933	Private room	90	7	0	NaN	NaN	1	23

48895 rows × 16 columns

Exploring & cleaning

Remove duplicates & unnecessary features

	index	neighbourhood_group	room_type	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365	price
0	0	Brooklyn	Private room	1	9	0.21	6	365	149
1	1	Manhattan	Entire home/apt	1	45	0.38	2	355	225
2	2	Manhattan	Private room	3	0	NaN	1	365	150
3	3	Brooklyn	Entire home/apt	1	270	4.64	1	194	89
4	4	Manhattan	Entire home/apt	10	9	0.10	1	0	80
...	...	...	...	...	...	...	...	...	...
45664	48890	Brooklyn	Private room	2	0	NaN	2	9	70
45665	48891	Brooklyn	Private room	4	0	NaN	2	36	40
45666	48892	Manhattan	Entire home/apt	10	0	NaN	1	27	115
45667	48893	Manhattan	Shared room	1	0	NaN	6	2	55
45668	48894	Manhattan	Private room	7	0	NaN	1	23	90

45669 rows × 9 columns

Exploring & cleaning

One-Hot encoding

reviews_per_month	calculated_host_listings_count	availability_365	neighbourhood_group_Bronx	neighbourhood_group_Brooklyn	neighbourhood_group_Manhattan	neighbourhood_group_Queens	neighbourhood_group_Staten Island	room_type_Entire home/apt	room_type_Private room	room_type_Shared room
0.21	6	365	0	1	0	0	0	0	1	0
0.38	2	355	0	0	1	0	0	1	0	0
NaN	1	365	0	0	1	0	0	0	1	0
4.64	1	194	0	1	0	0	0	1	0	0
0.10	1	0	0	0	1	0	0	1	0	0
...	...	...	...	...	...	...	...	...	...	...
NaN	2	9	0	1	0	0	0	0	1	0
NaN	2	36	0	1	0	0	0	0	1	0
NaN	1	27	0	0	1	0	0	1	0	0
NaN	6	2	0	0	1	0	0	0	0	1
NaN	1	23	0	0	1	0	0	0	1	0

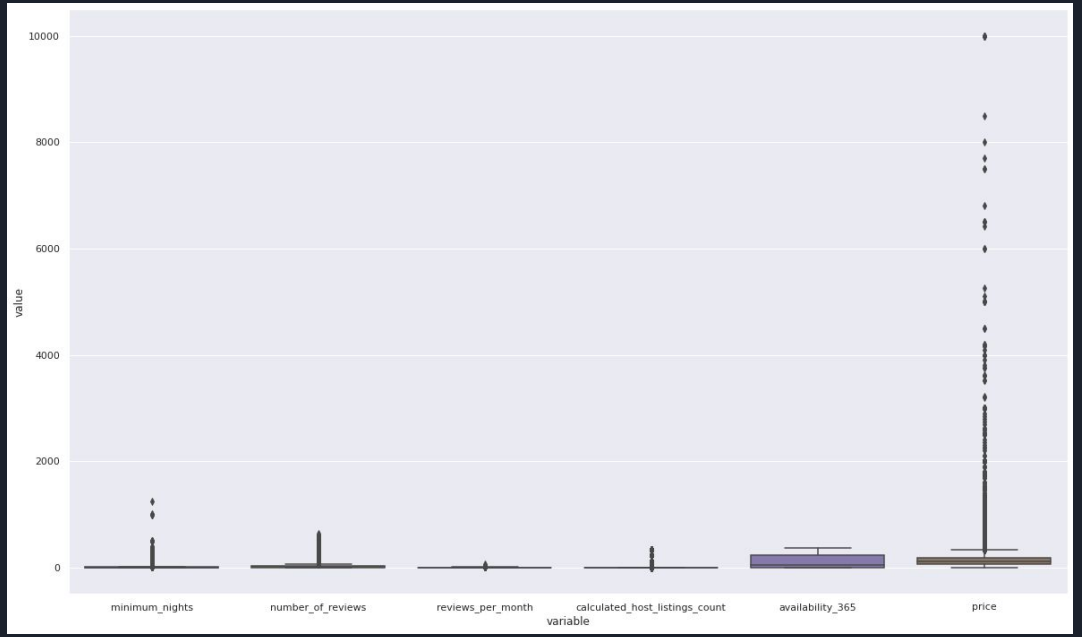
Transforming all categorical features to numeric

Exploring & cleaning

Extreme values

Let extreme values be the quantities that are bigger than the upper limit of a box plot, i.e, values bigger than $Q_3 + 1.5\ (Q_3 - Q_1)$

	feature	upper limit	percentage of extreme values
0	minimum_nights	11.0	13.58
1	number_of_reviews	58.5	12.31
2	calculated_host_listings_count	3.5	14.48
3	availability_365	567.5	0.00
4	price	334.0	6.09



Exploring & cleaning

Handling missing values & data scaling

Split data to train, validation and test datasets

We use knn regressor algorithm to predict the missing values depending on 5 nearest neighbours

The distributions of the numeric features are not normally distributed and skewed

We use Min-Max scaling to scale our data

Now, since we cleaned our data and no longer have missing values we can proceed to build our QuAMs

Evaluation and QuAM Development

Classification

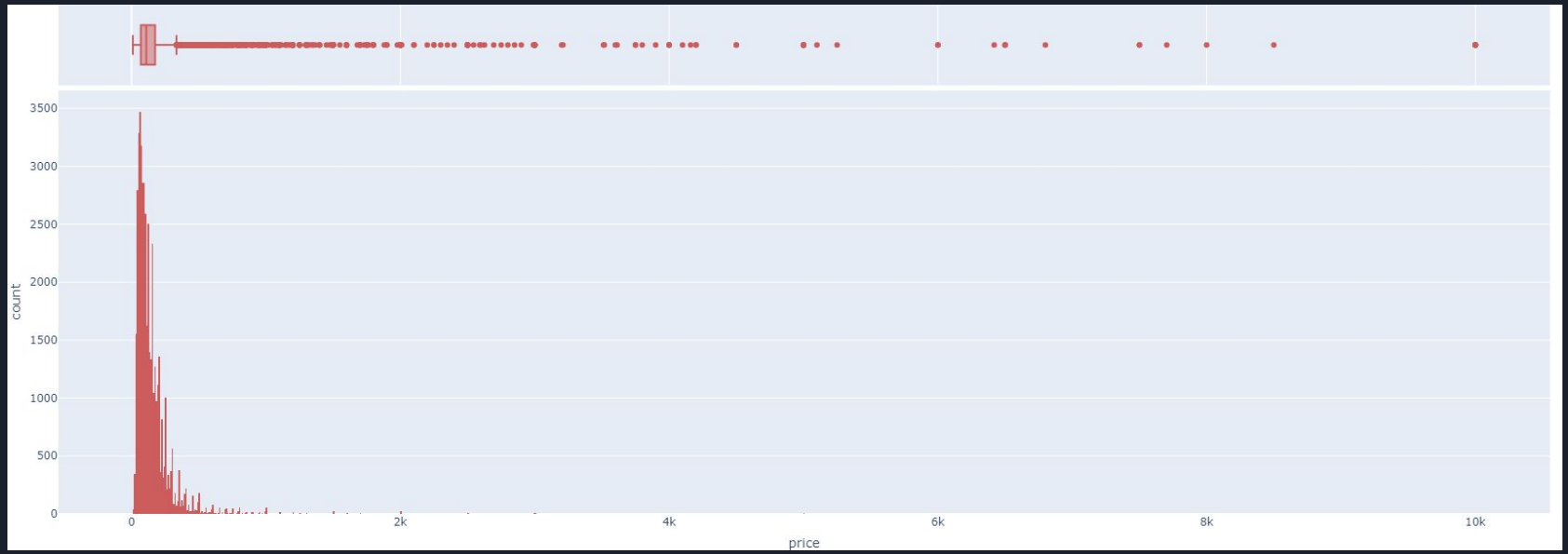
We want to classify our target variable (price) into 4 classes:

Cheap: prices $< Q_1$

Medium: $Q_1 \leqslant$ prices $< Q_3$

Expensive: $Q_3 \leqslant$ prices $< Q_3 + 1.5\,(Q_3 - Q_1)$

Very Expensive: $Q_3 + 1.5\,(Q_3 - Q_1) \leqslant$ prices



Classification

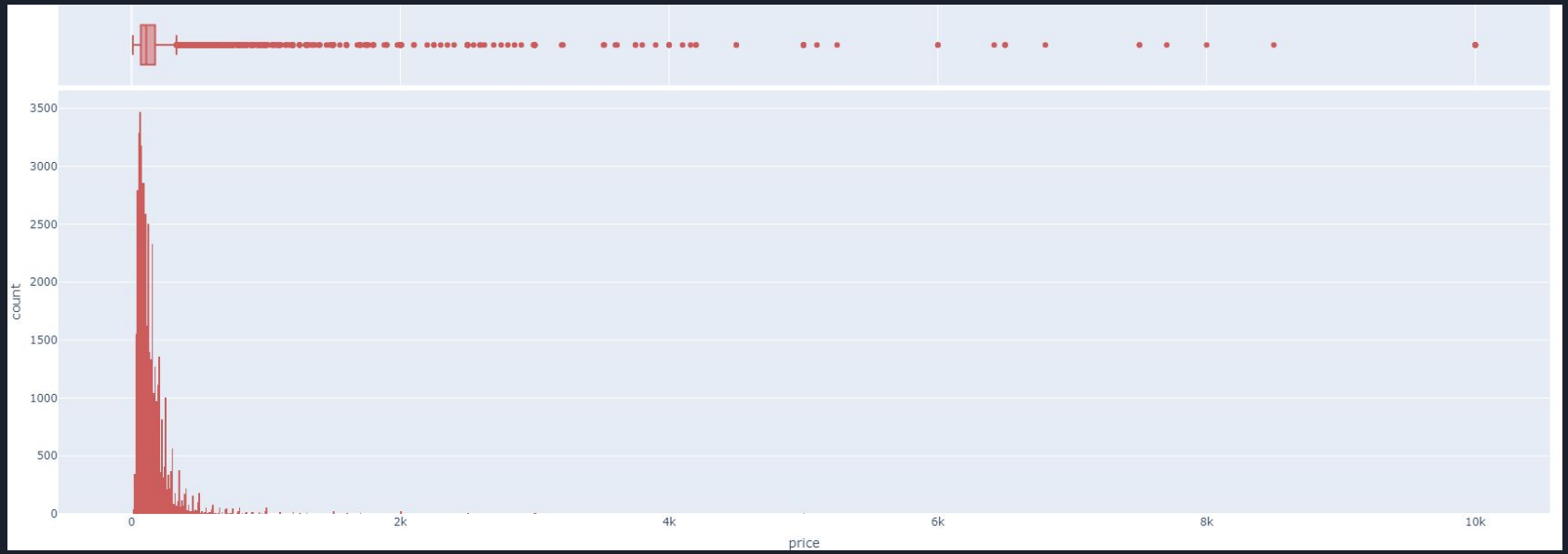
In our case:

Cheap: prices < 70

Medium: 70 ⩽ prices < 175

Expensive: 175 ⩽ prices < 334

Very Expensive: 334 ⩽ prices



Classification

In our case:

Cheap: prices: < 70

Medium: 70 ⩽ pri < 1

Expensive: 175 ⩽ prices < 334

ery Expe sive: 334 ⩽ rices

Unrealistic

Classification

Another approach …

Private room

Shared room

Entire home/apt

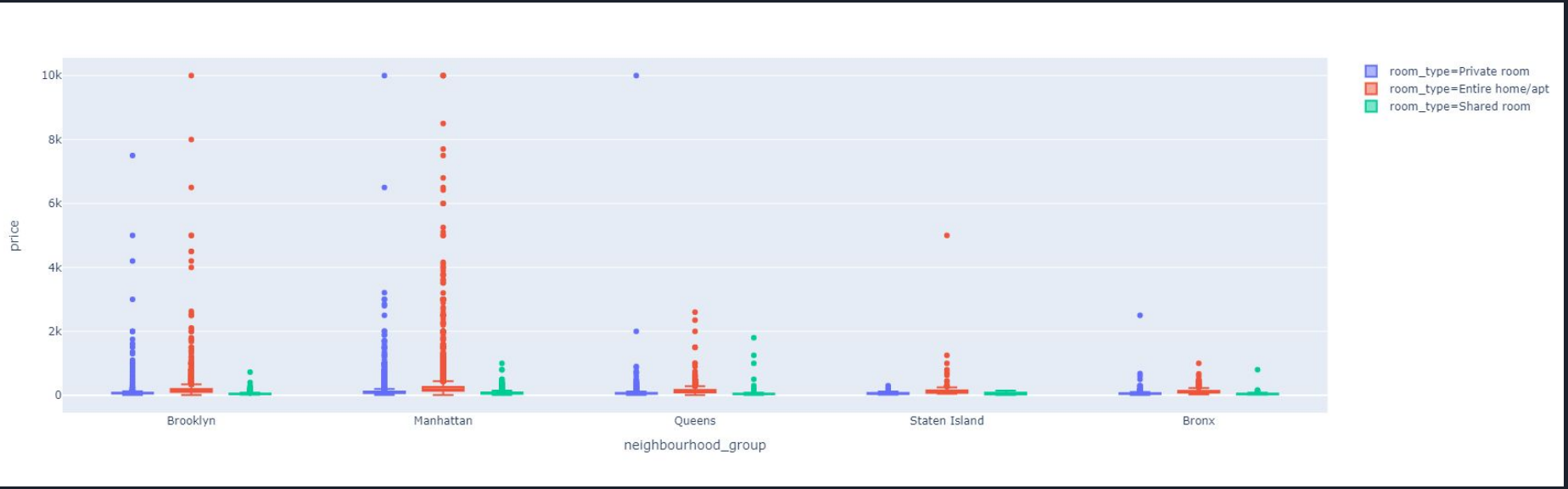
https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City

Each neighbourhood has different valuations for listings

Each room type has a different rental value

Classification

By-market-classification



Classification

By-market-classification

This approach considers each market separately

Extreme values for each market is different

Thresholds are [Q_1 , Q_3 , $Q_3 + 1.5\,(Q_3 - Q_1)$]

We will work with classification by market and in general for comparison reasons

	Market	Thresholds
0	Brooklyn_Private room	[50.0, 80.0, 125.0]
1	Brooklyn_Entire home/apt	[105.0, 199.0, 340.0]
2	Brooklyn_Shared room	[30.0, 49.5, 78.75]
3	Manhattan_Private room	[67.0, 120.0, 199.5]
4	Manhattan_Entire home/apt	[140.0, 260.0, 440.0]
5	Manhattan_Shared room	[49.0, 90.0, 151.5]
6	Queens_Private room	[47.0, 75.0, 117.0]
7	Queens_Entire home/apt	[90.0, 169.0, 287.5]
8	Queens_Shared room	[30.0, 55.0, 92.5]
9	Staten Island_Private room	[40.0, 75.0, 127.5]
10	Staten Island_Entire home/apt	[75.0, 151.25, 265.625]
11	Staten Island_Shared room	[29.0, 75.0, 144.0]
12	Bronx_Private room	[40.0, 70.0, 115.0]
13	Bronx_Entire home/apt	[80.0, 140.0, 230.0]
14	Bronx_Shared room	[28.0, 55.5, 96.75]

Classification

Decision Tree

Classification

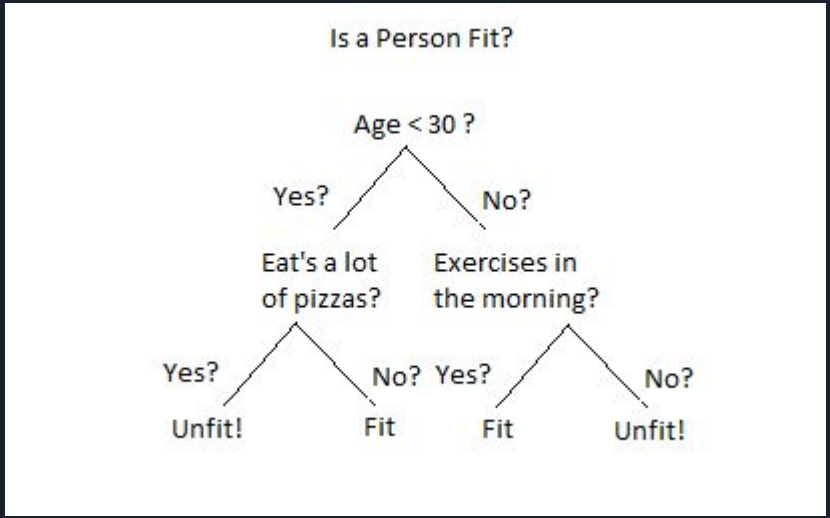
Decision Tree - Quick overview

predictive modelling approach

go from observations about an item to conclusions about the item's target value

leaves represent class labels

hyper parameter: maximum depth

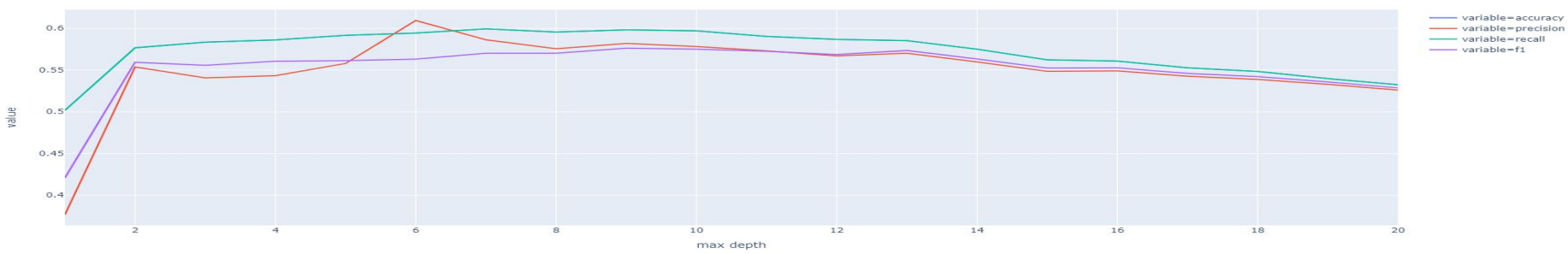


https://chiragsehra42.medium.com/decision-trees-explained-easily-28f23241248

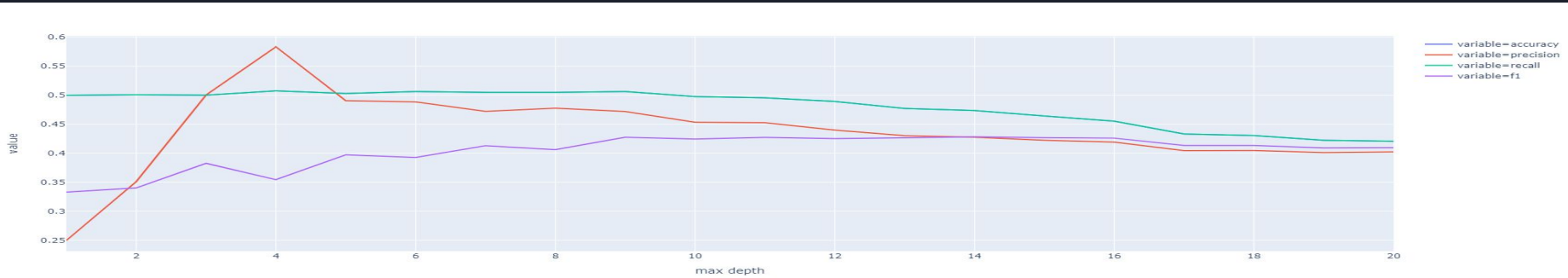
Classification

Decision Tree - score metrics

General classification



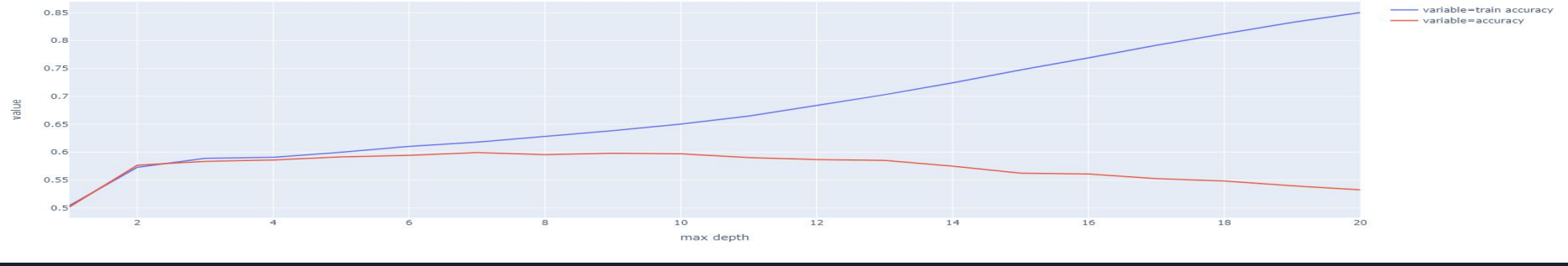
By market classification



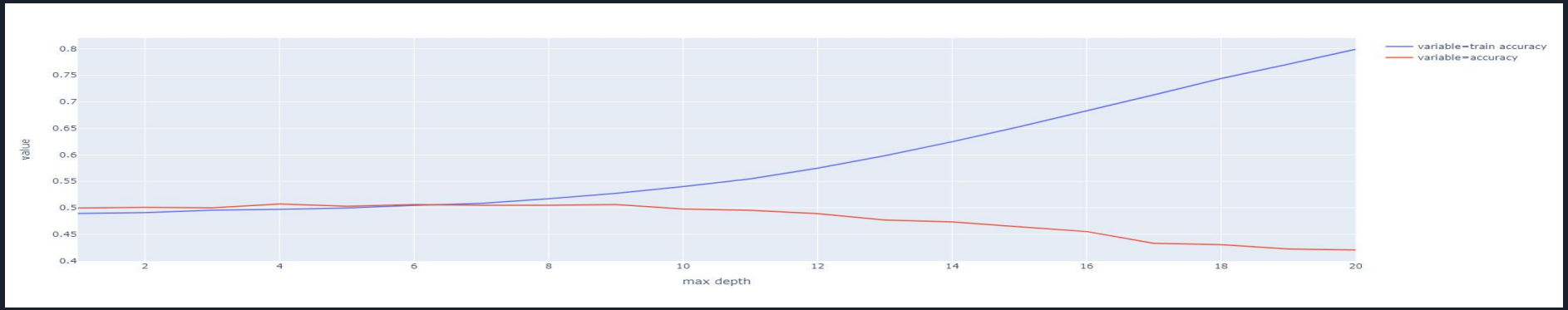
Classification

Decision Tree - sensitivity to hyperparameter

General classification

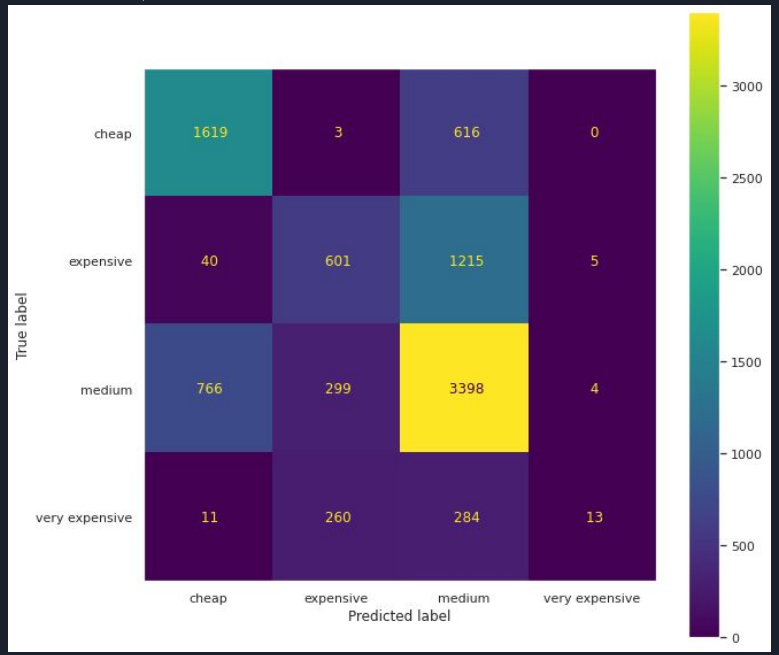


By market classification

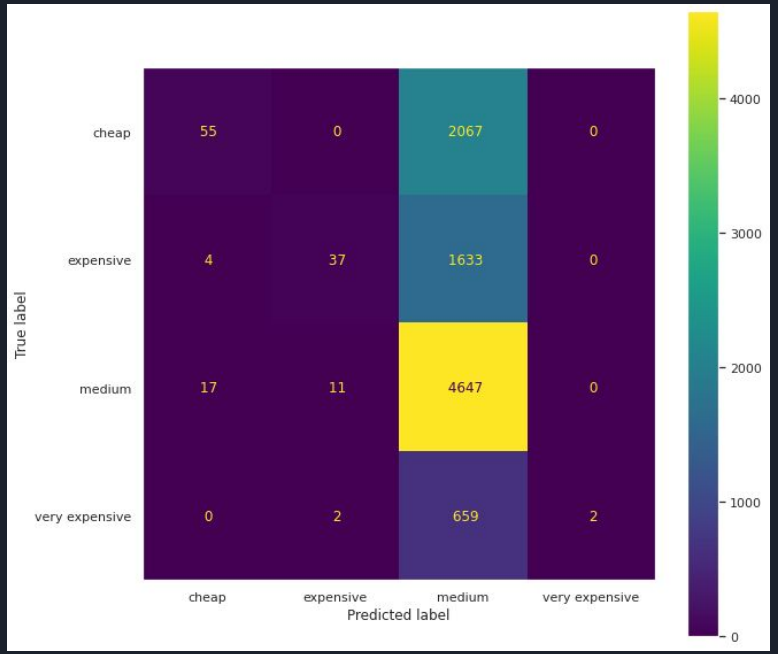


Classification

Decision Tree - Correct classifications and misclassifications



General classification optimum max_depth = 7



By market classification optimum max_depth = 4

Classification

Decision Tree

For the optimal max_depth in each classification approach, the general data performed better than by market Classification

Accuracy for general is:

Train 0.618 Test 0.6165

Accuracy for market is:

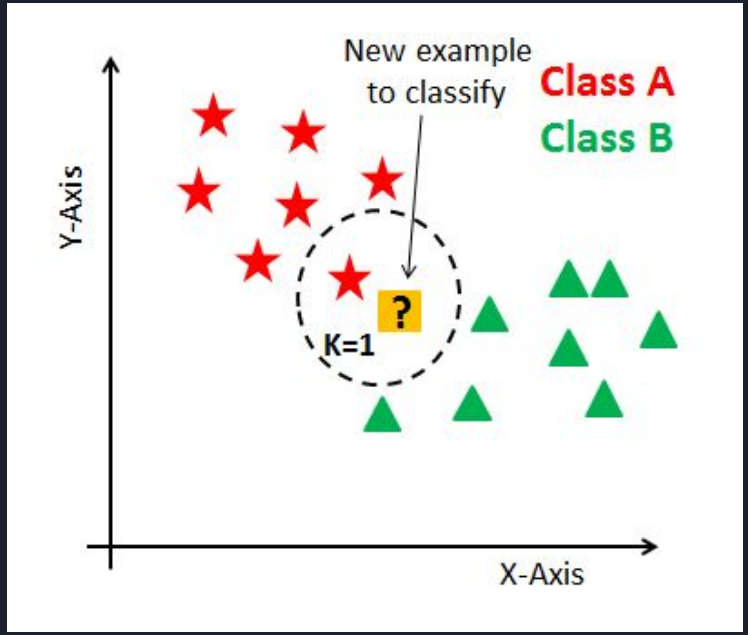
Train 0.4972 Test 0.519

Classification

k-Nearest Neighbours

QuAM - Nearest Neighbours KNN

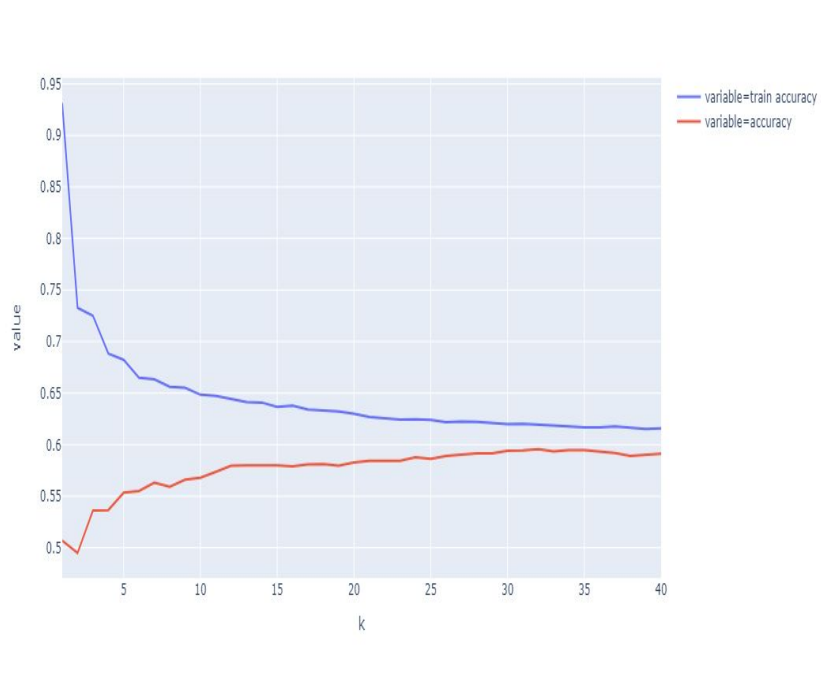
- Classification based QUAM
- Evaluation based on distance
- Major vote



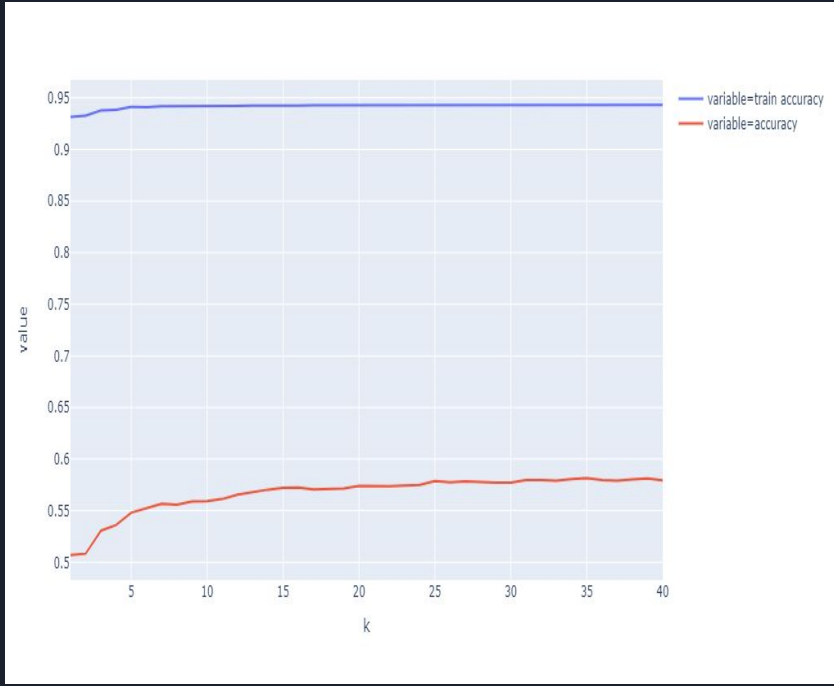
https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn

QuAM - Nearest Neighbours KNN

Hyper parameter sensitivity - Accuracy



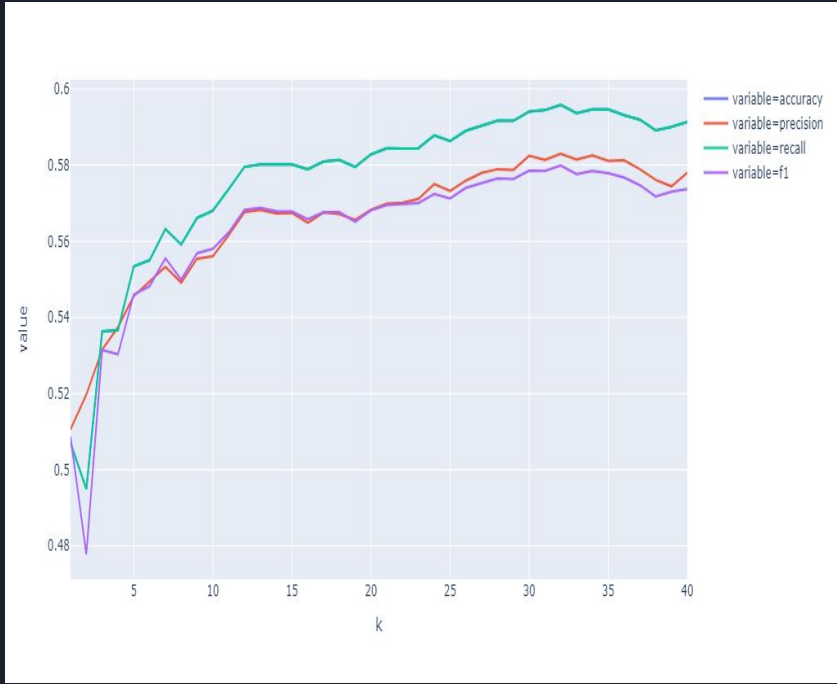
K- sensitive weighted uniformly



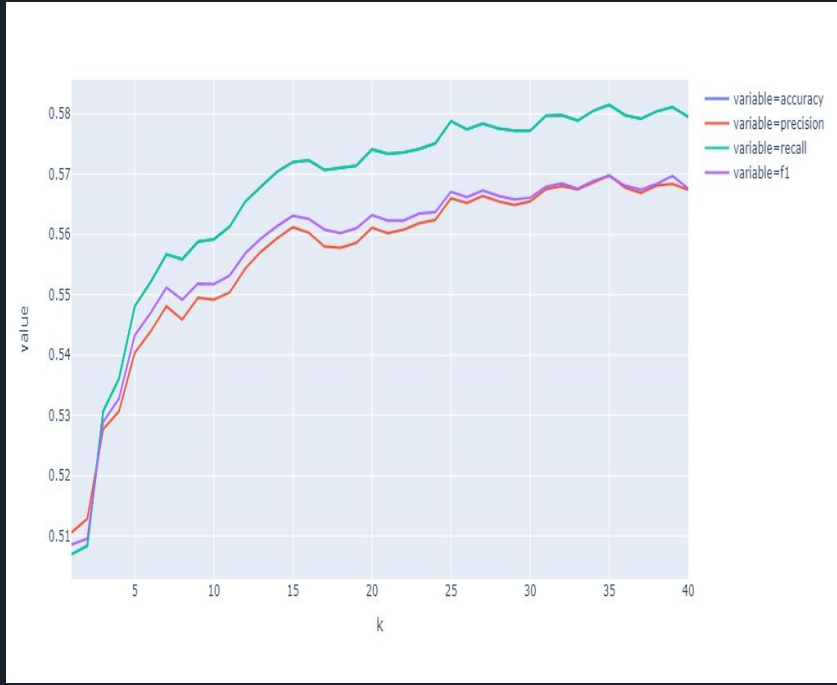
K- sensitive weighted with distance

QuAM - Nearest Neighbours KNN

Hyper parameter sensitivity - Accuracy, Precision, Recall and F1



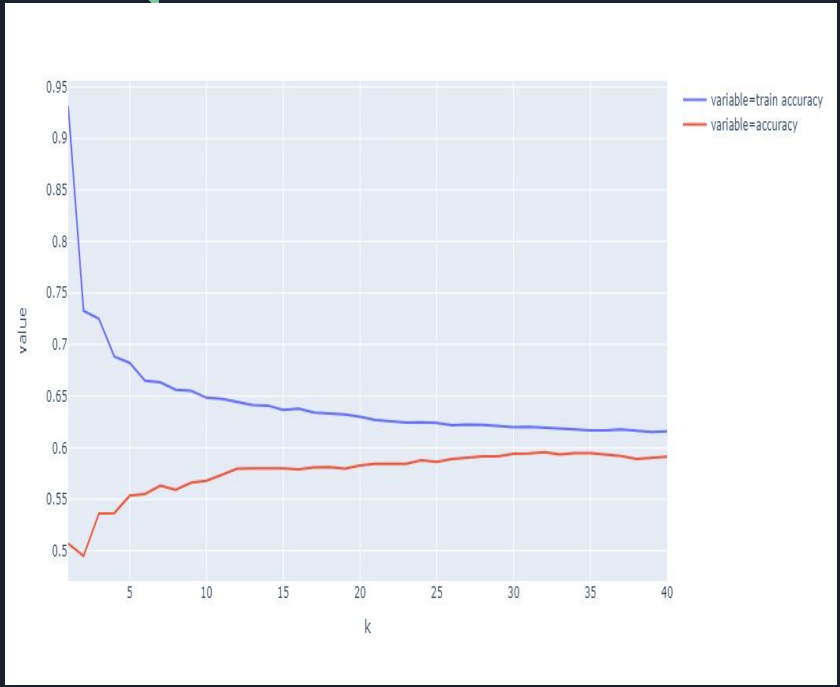
K- sensitive with uniform weights



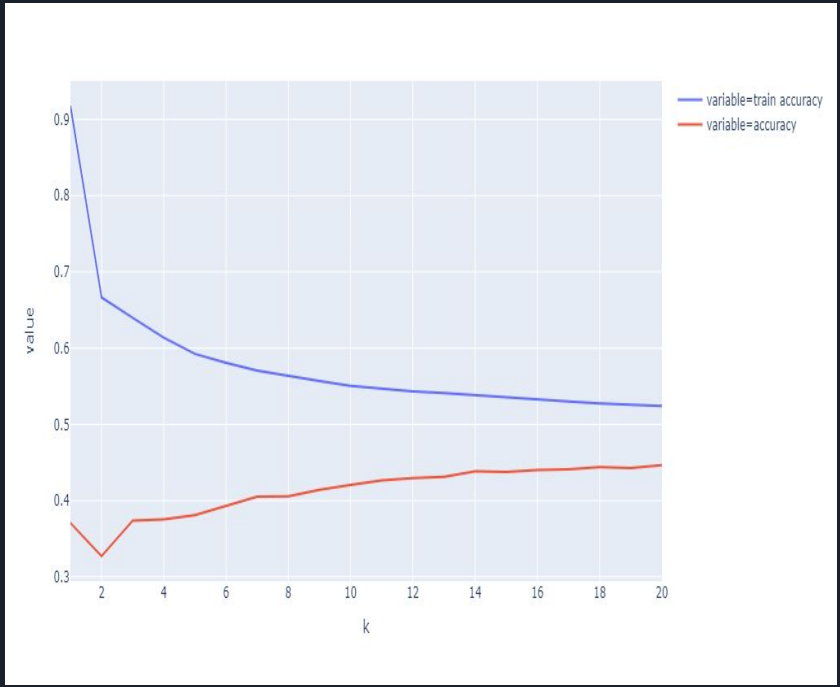
K- sensitive with uniform weights

QuAM - Nearest Neighbours KNN

Hyper parameter sensitivity - Accuracy for all data and markets specific



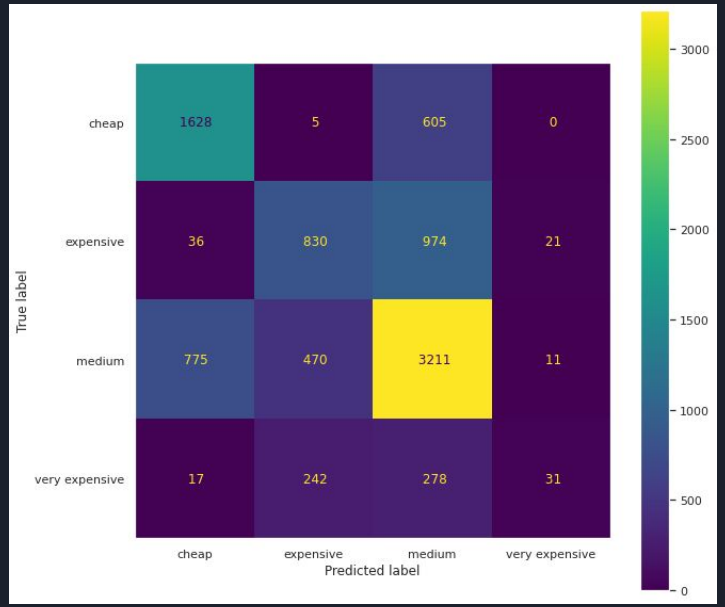
General data (optimum K=33 with accuracy)



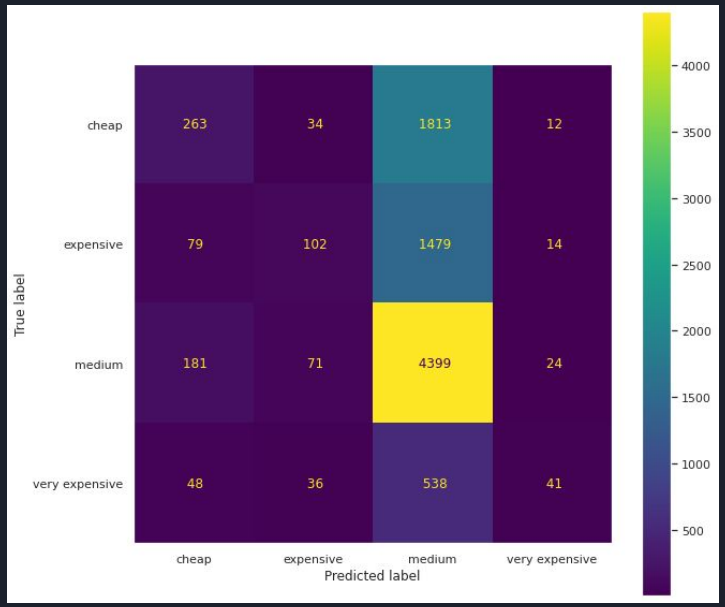
Market specific (optimum K=33)

QuAM - Nearest Neighbours KNN

Hyper parameter sensitivity - Accuracy for all data and markets specific



General data (optimum K=33)



Market specific (optimum K=33)

Classification

KNN

Optimum K for both is 33

Accuracy for general is:

Train 0.6191 Test 0.6018

Accuracy for market is:

Train 0.5137 Test 0.5

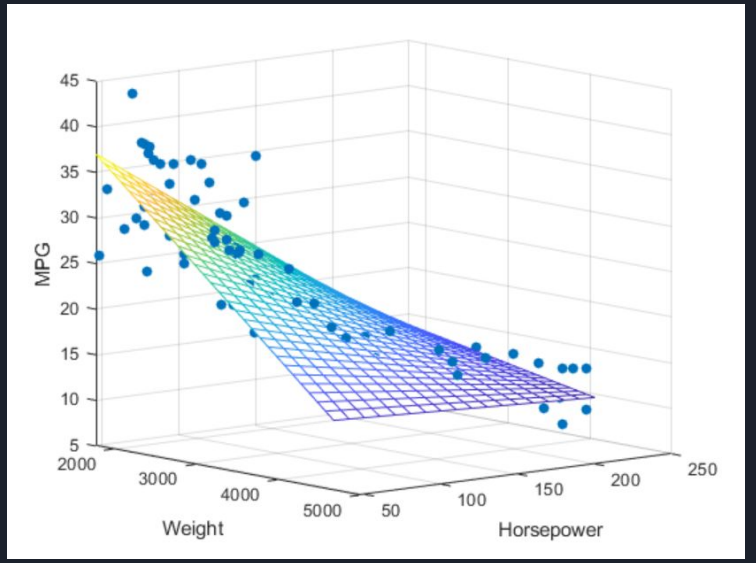
Linear Regression

- Not a classification QUAM
- Generalizes the data with a function
- For a univariate data

$$\hat{y} = m * x + b + \epsilon$$

- Applicable to multi variate

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n + \epsilon$$



https://meraju.com/linear-regression/

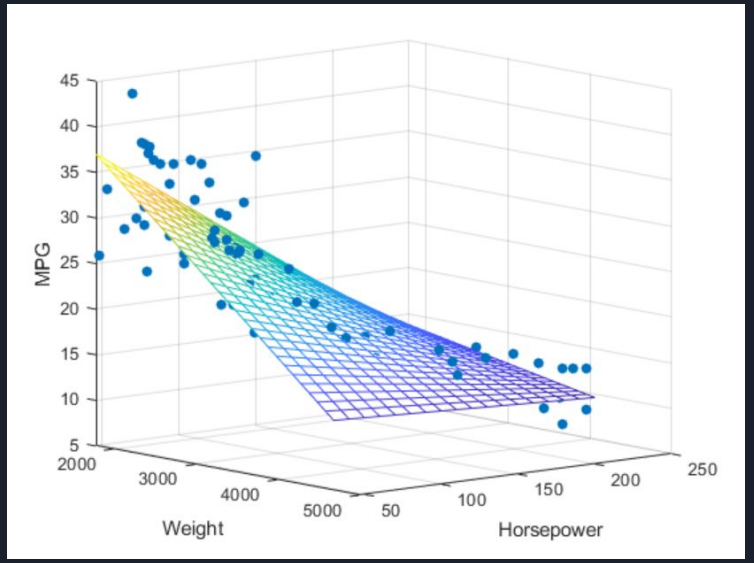
Linear Regression

- Evaluation metrics

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$



https://meraju.com/linear-regression/

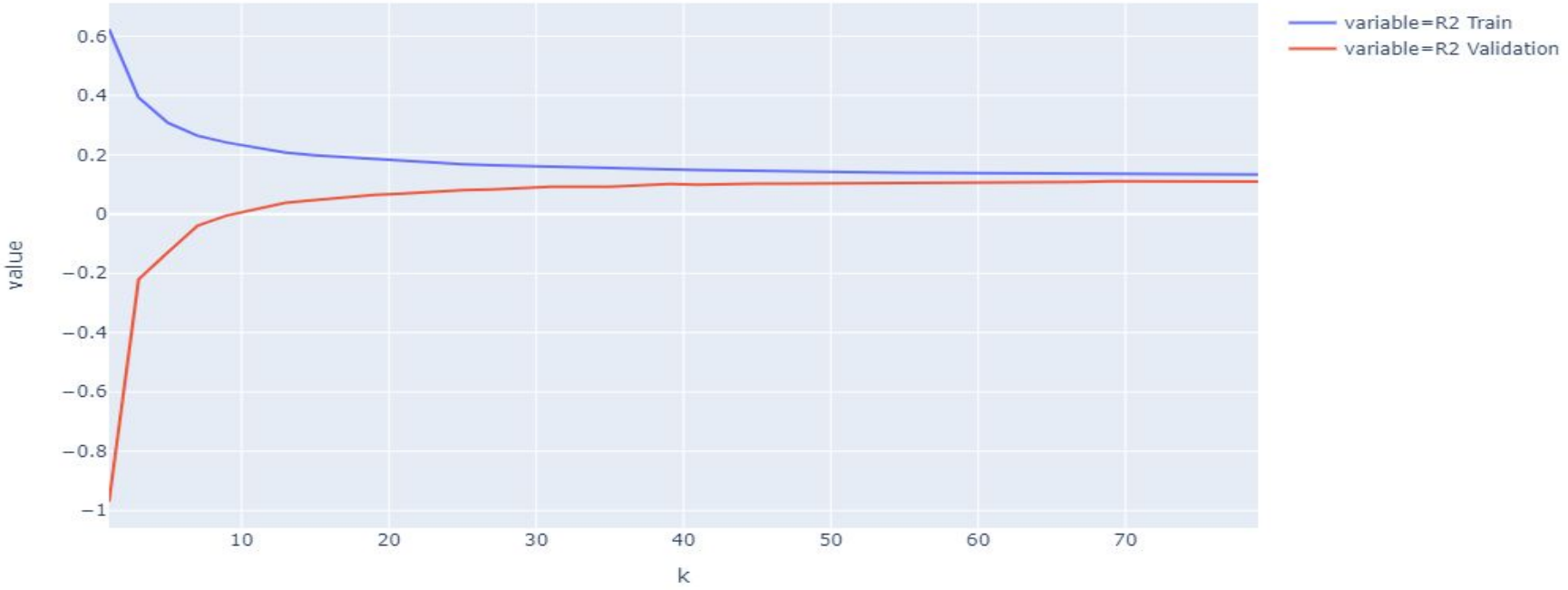
Linear Regression - Evaluation metrics

	Training Data	Test Data
MAE	77.6577	75.7711
MSE	58903.0359	50694.8259
R squared	8.68	9.02

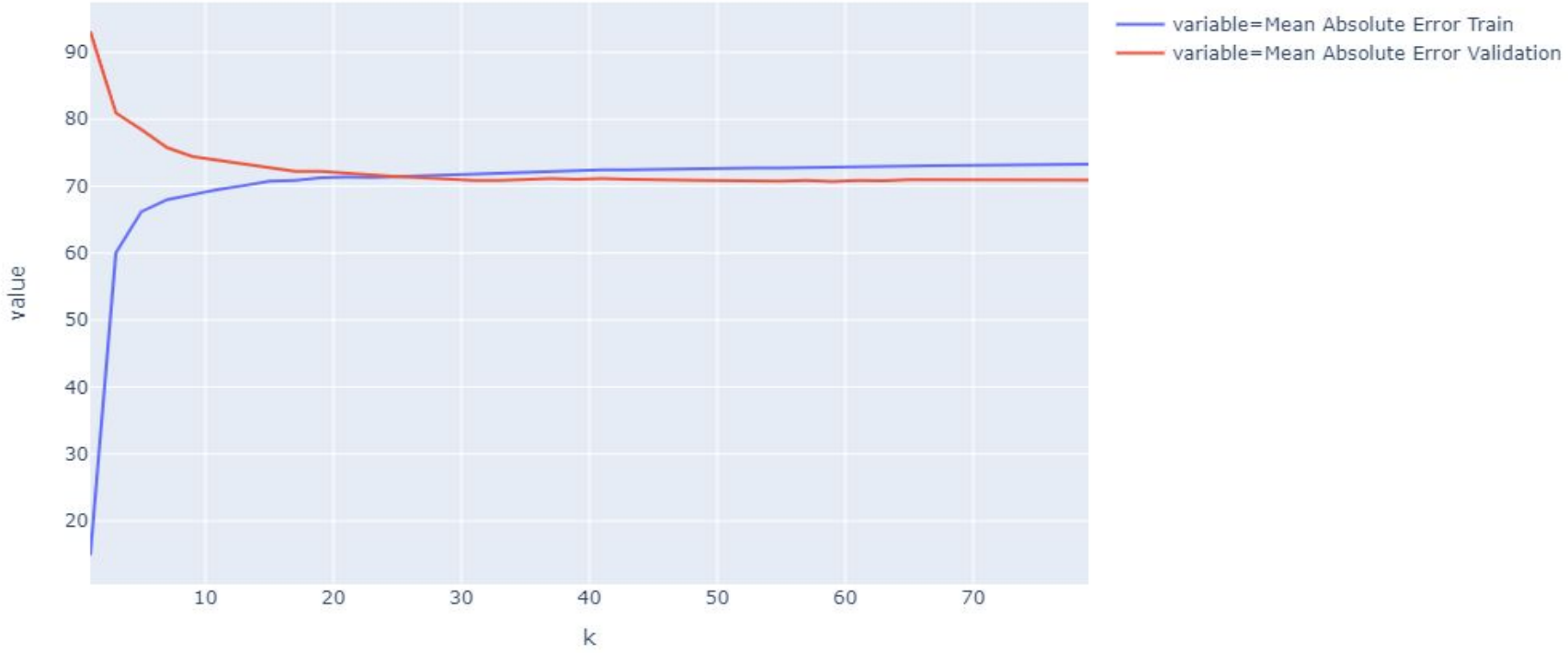
KNN Regression - Definition

- Not a classification method
- Averages the observation in the neighborhood to get predict continuous value
- Hyperparameter: K neighbours

KNN Regression - Results



KNN Regression - Results



KNN Regression - Results

- Best K based on comparison by R-Squared and MAE is 23

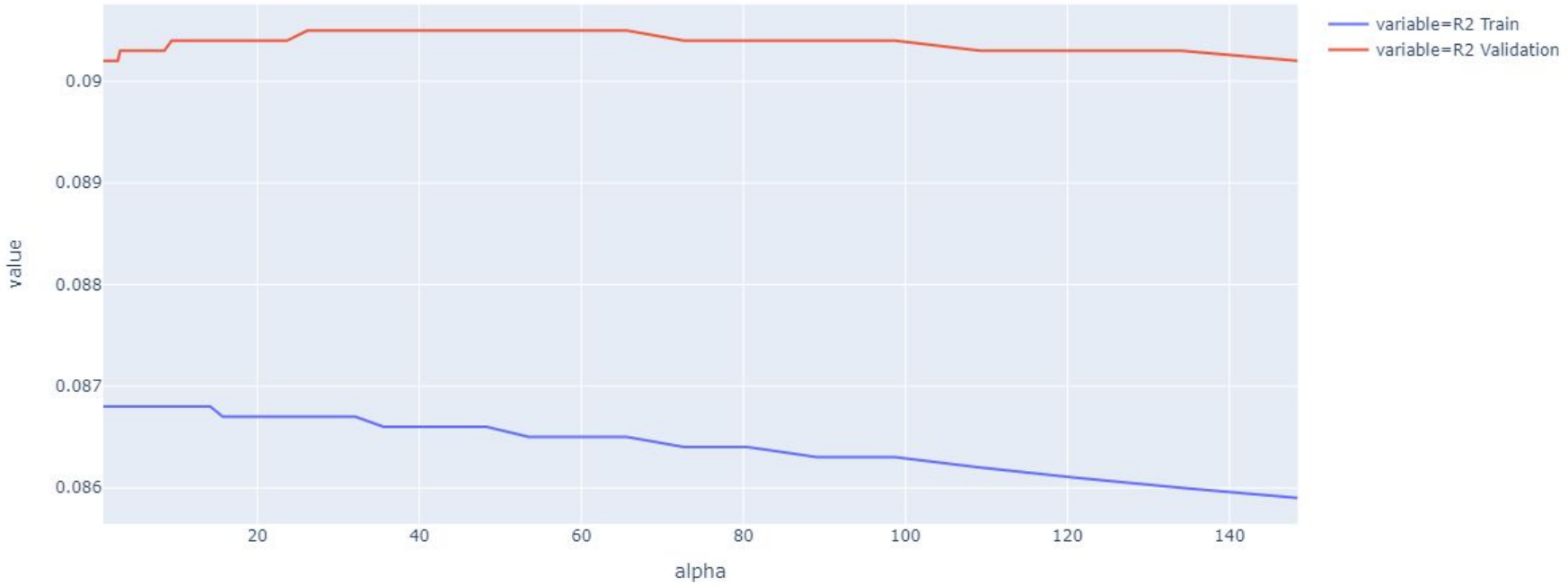
- R2 Train: 13.57
- R2 Test: 12.68

Ridge Regression - Definition

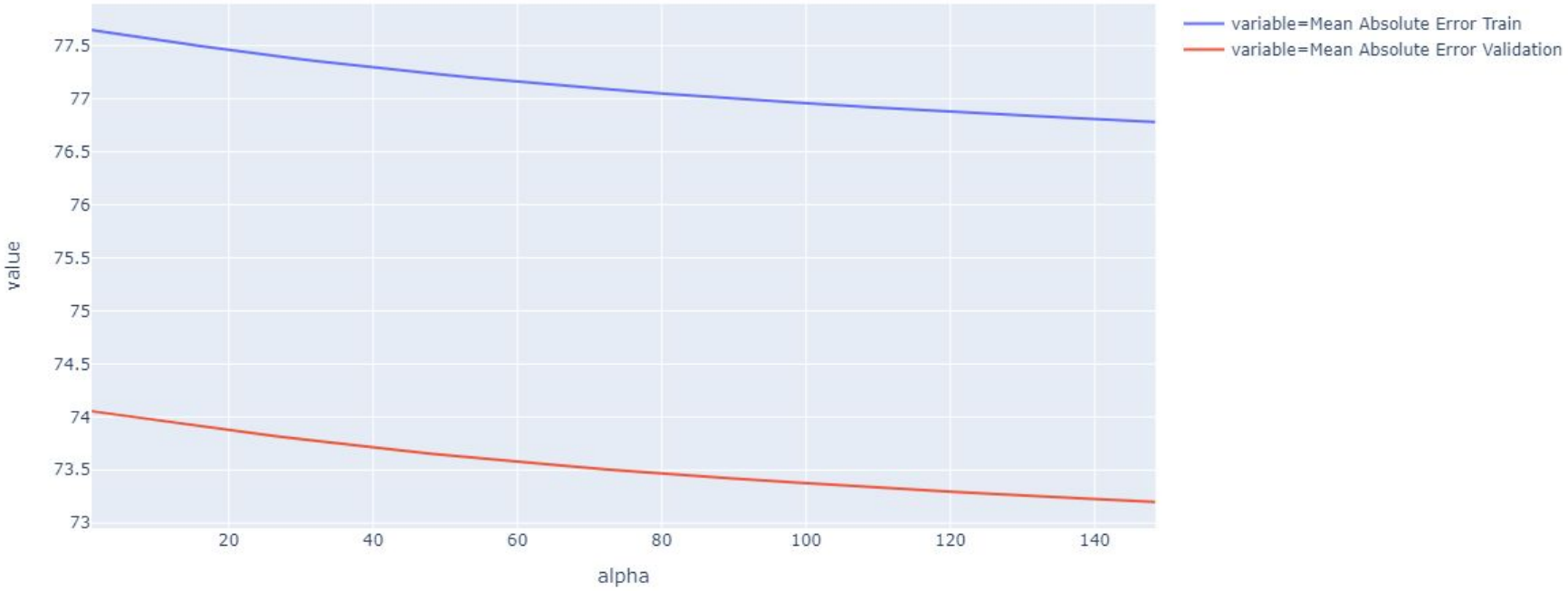
- Linear least squares with l2 regularization.
- $$\sum_{i=1}^{M} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{p} w_j \times x_{ij} \right)^2 + \lambda \sum_{j=0}^{p} w_j^2$$
- Hyperparameter: alpha - Regularization Strength

https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b

Ridge Regression - Results



Ridge Regression - Results



Ridge Regression - Results

- Best alpha based on comparison by R-Squared: 26.18

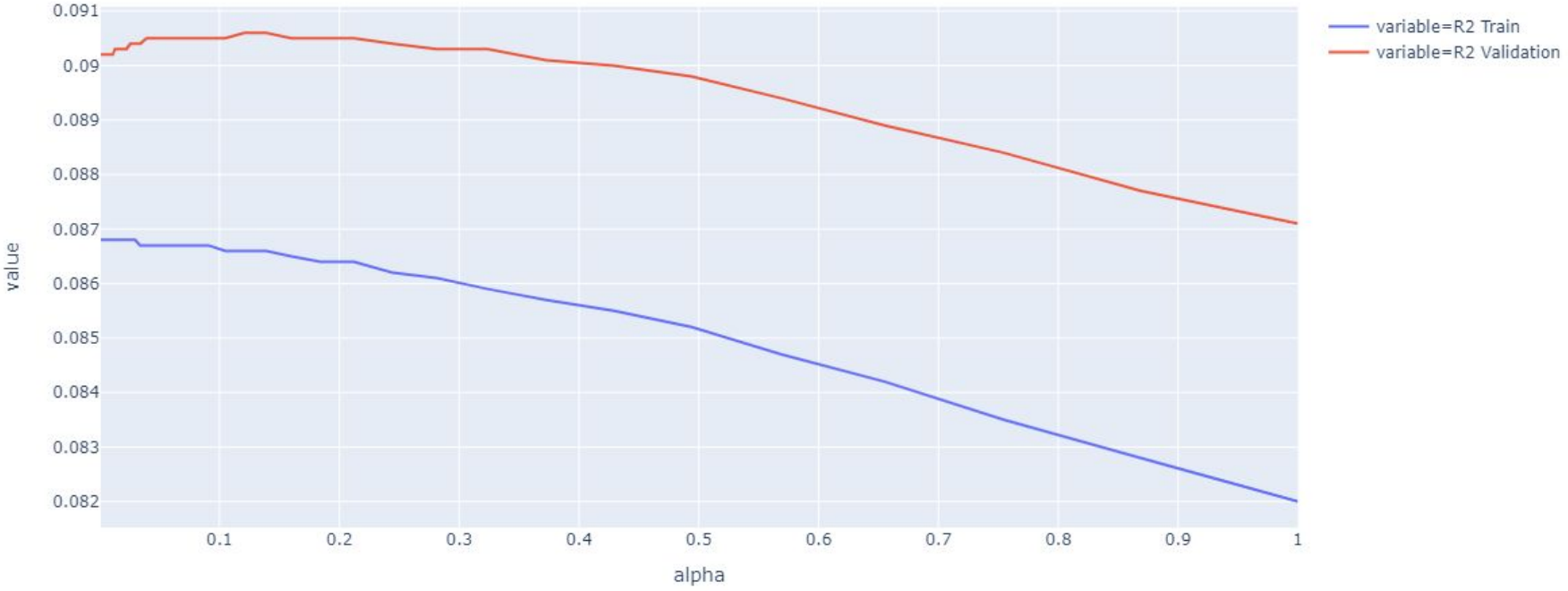
- R2 Train: 8.67
- R2 Test: 9.06

Lasso Regression - Definition

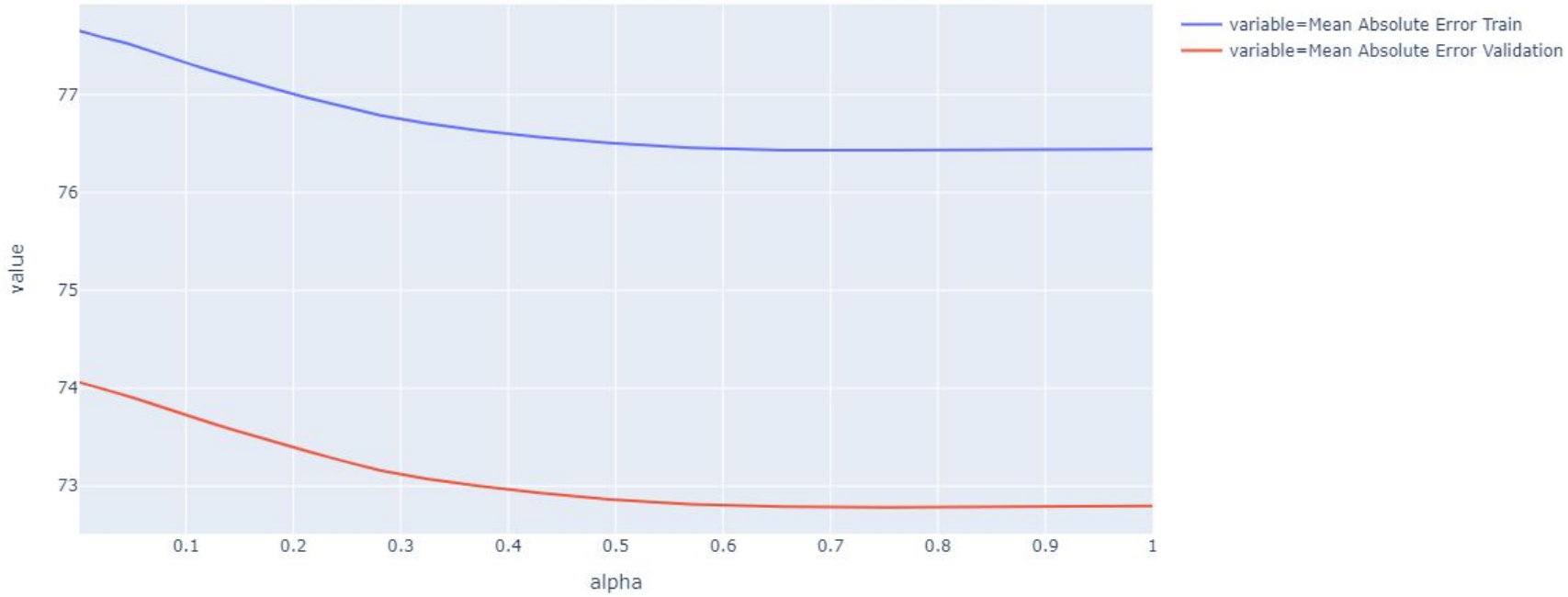
- Linear Model trained with L1 prior as regularizer.
- $$\sum_{i=1}^{M} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{p} w_j \times x_{ij} \right)^2 + \lambda \sum_{j=0}^{p} |w_j|$$
- Hyperparameter: alpha - Constant that multiplies the L1 term.

https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b

Lasso Regression - Results



Lasso Regression - Results



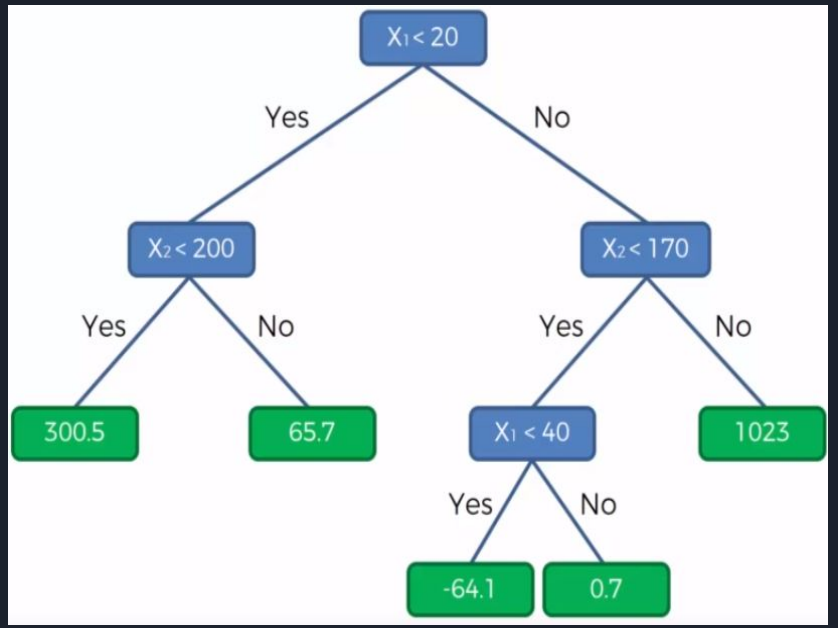
Lasso Regression - Results

- Best alpha based on comparison by R-Squared: 0.12

- R2 Train: 8.66
- R2 Test: 9.09

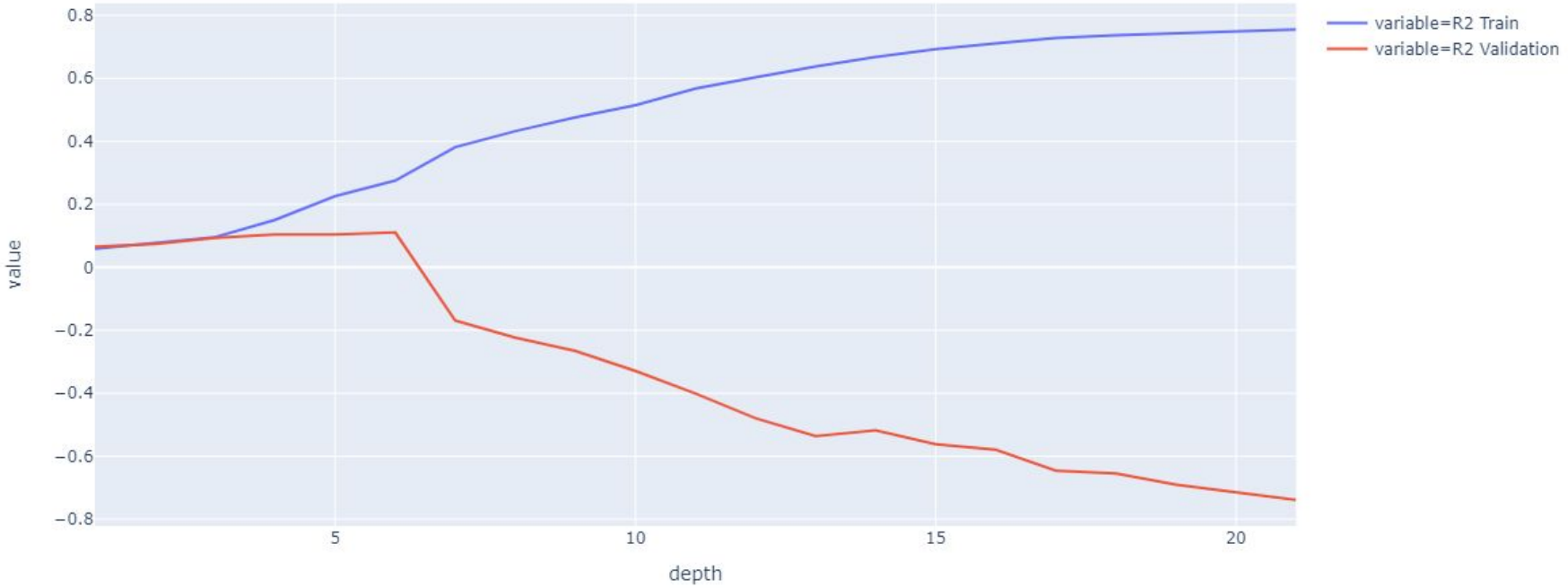
Decision Tree Regression - Definition

- Hyperparameter:
  Maximum Depth

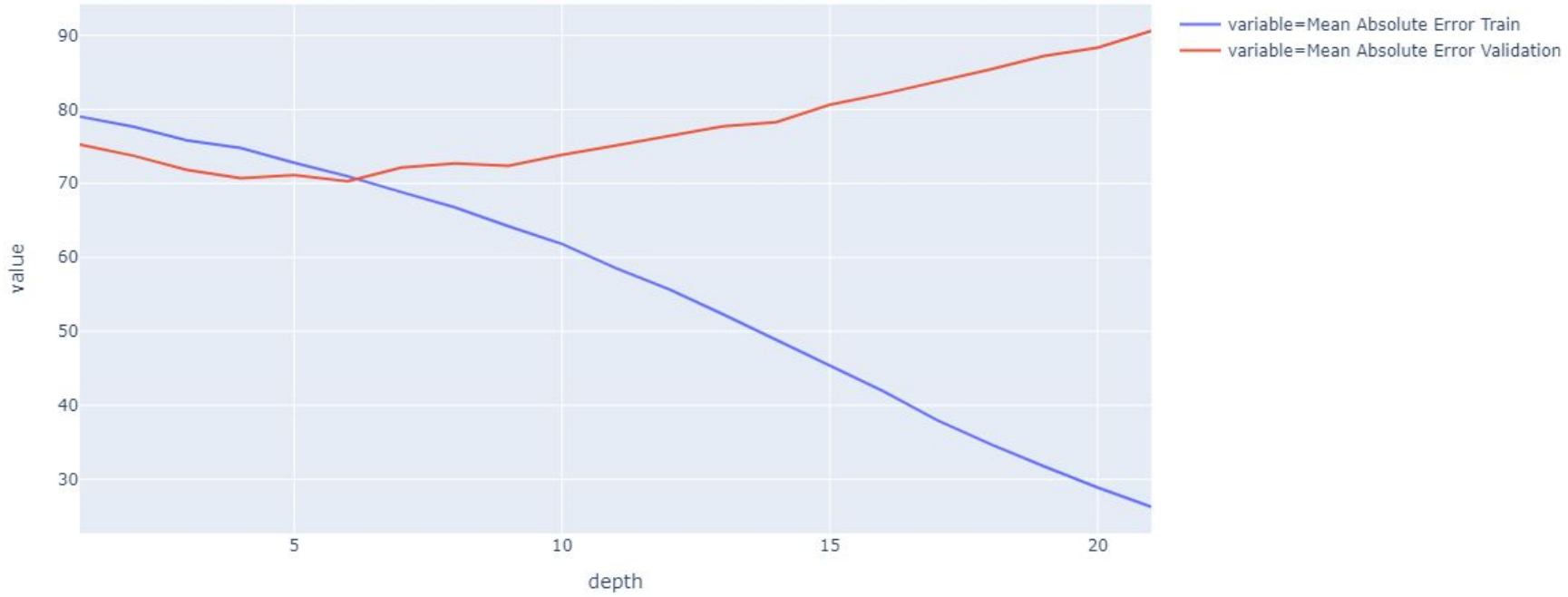


https://medium.com/pursuitnotes/decision-tree-regression-in-6-steps-with-python-1a1c5aa2ee16

Decision Tree Regression - Results



Decision Tree Regression - Results



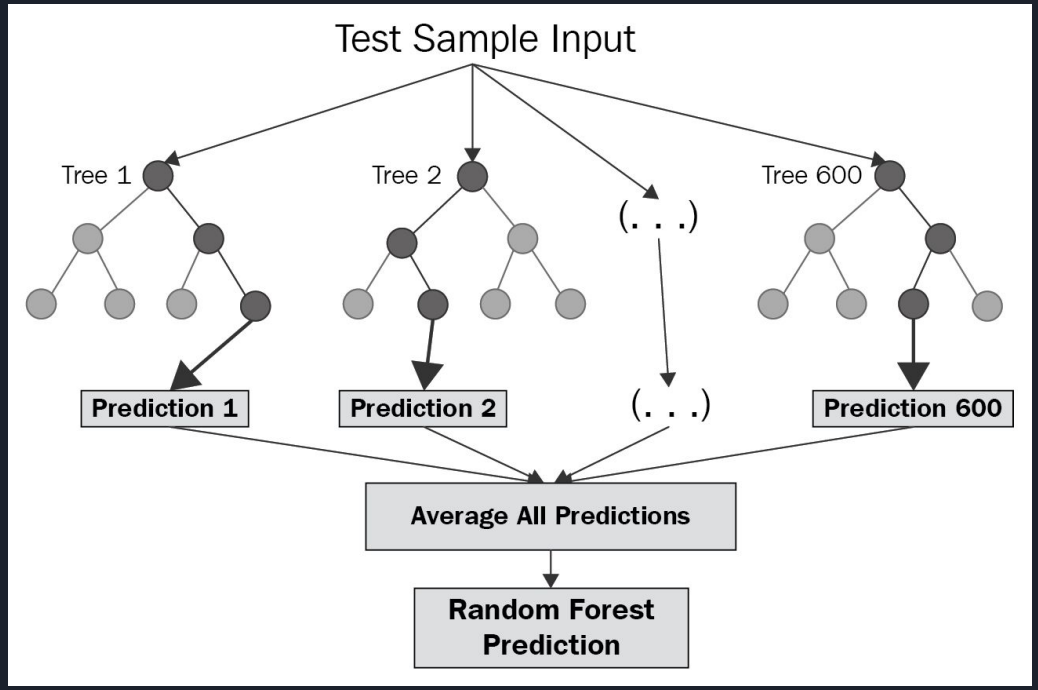
Decision Tree Regression - Results

- Best max depth based on comparison by all metrics: 6

- R2 Train: 27.52
- R2 Test: 18.03

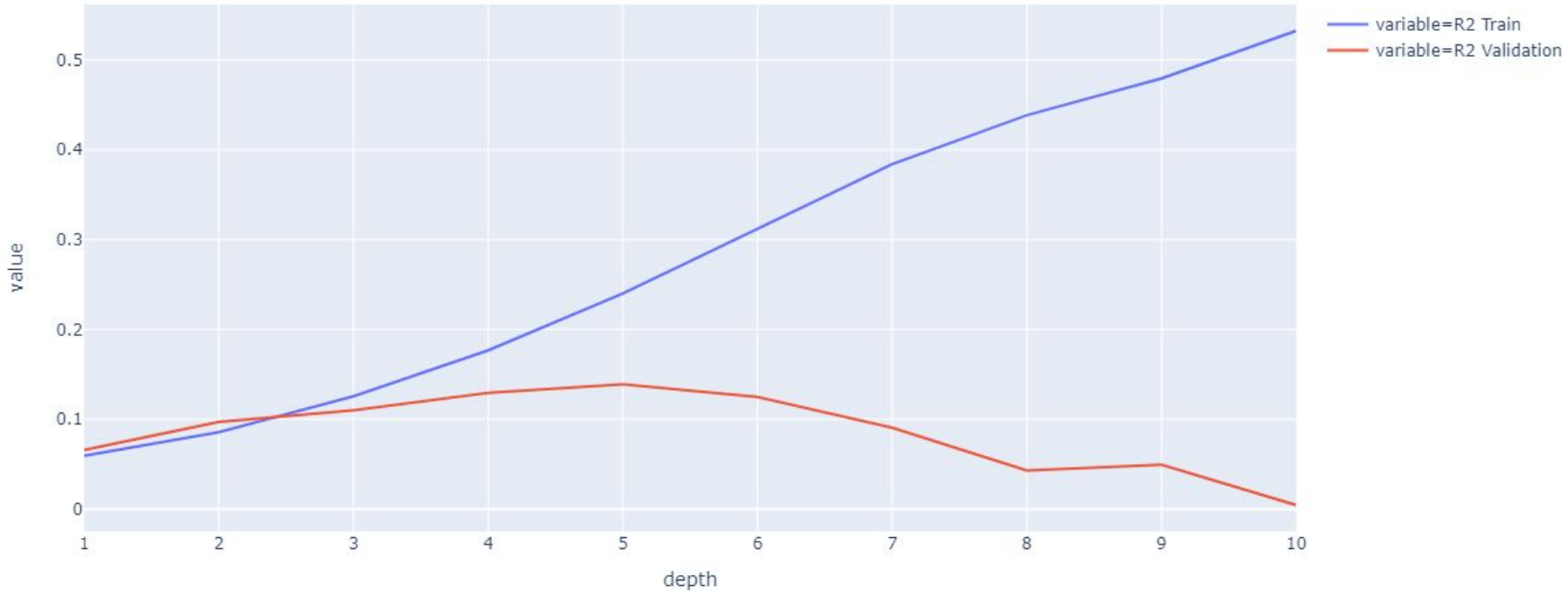
Random Forest Regression - Definition

- Ensemble method.
- Hyperparameter:
  Maximum Depth

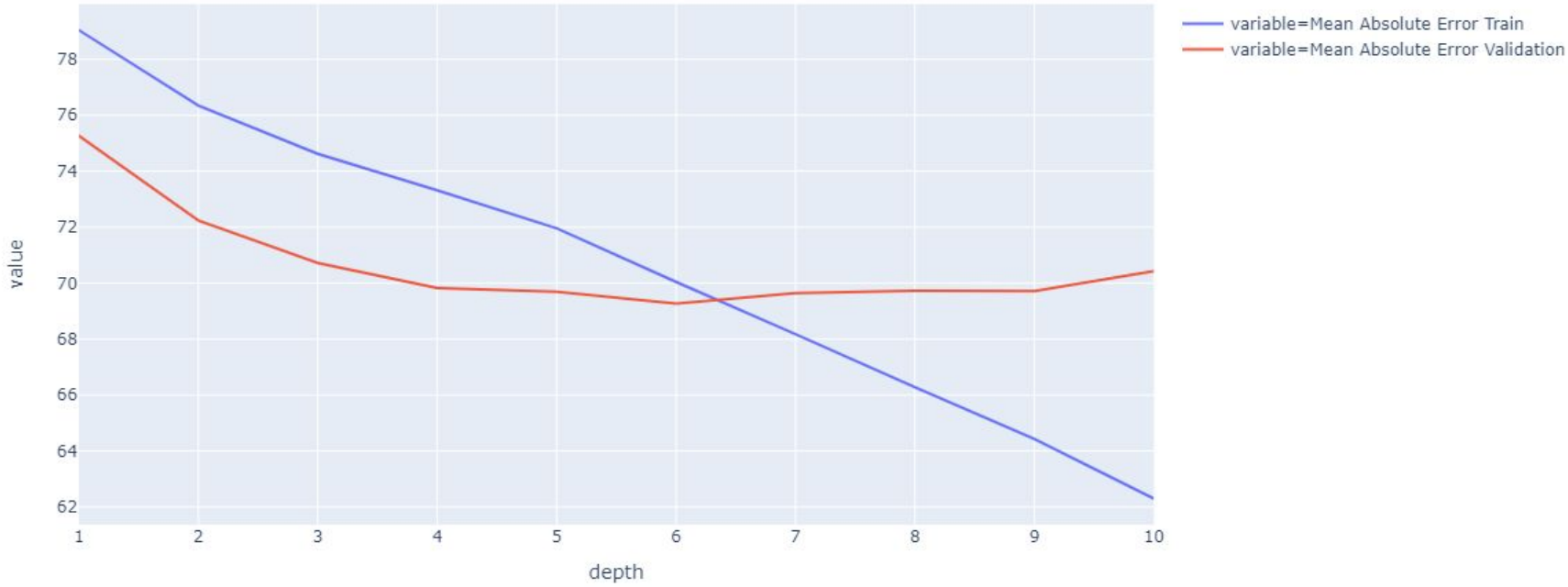


https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f

Random Forest Regression - Results



Random Forest Regression - Results



Random Forest Regression - Results

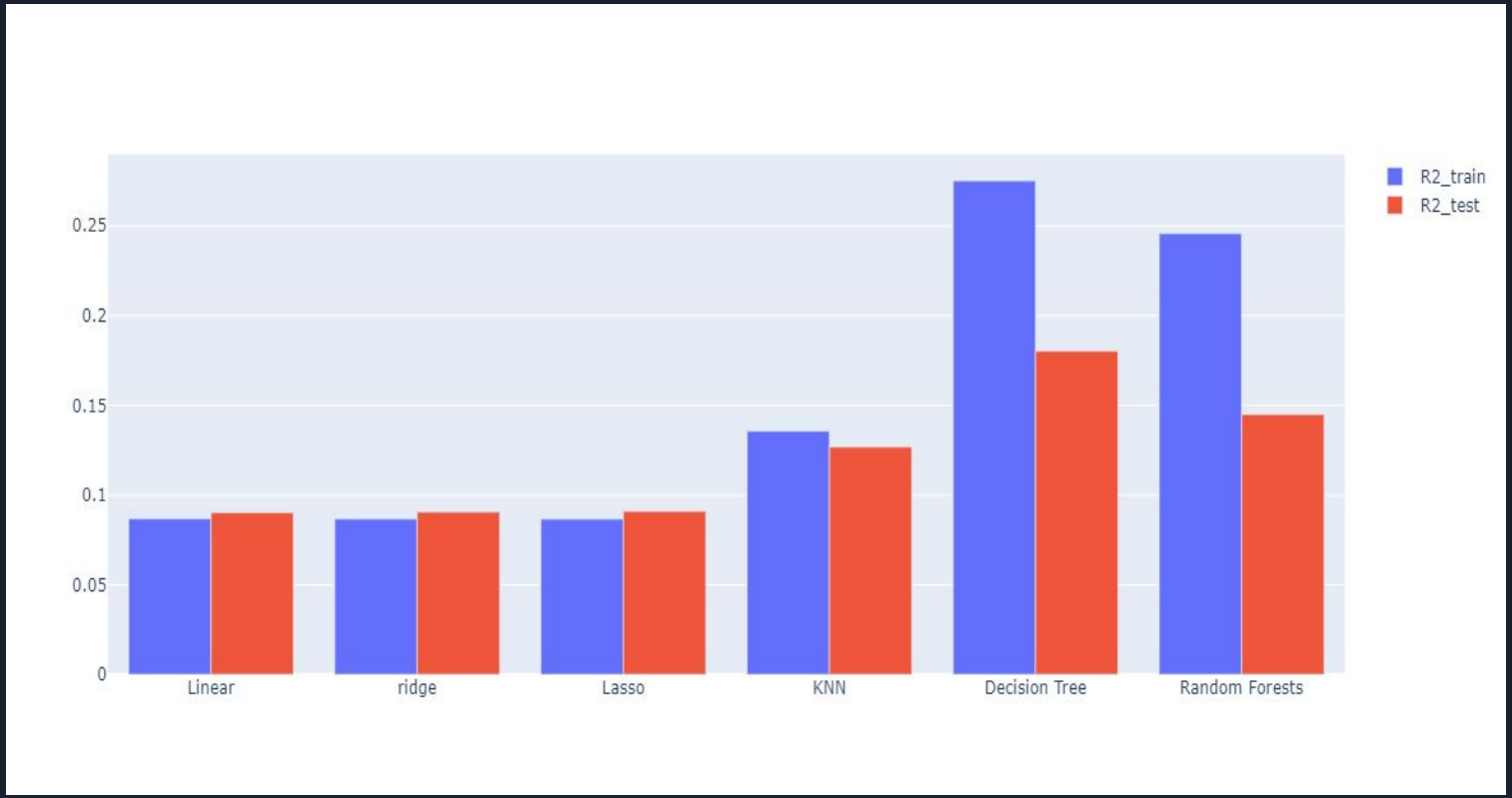
- Best max depth based on comparison by R2: 5

- R2 Train: 24.58
- R2 Test: 14.49

Regression Comparison

	R-squared Train %	R-squared Test %
Linear	8.68	9.02
Ridge	8.67	9.06
Lasso	8.66	9.09
KNN	***13.57***	***12.68***
Decision Tree	27.52	***18.03***
Random Forest	24.58	***14.49***

Regression Comparison



Conclusion

- Low values for regression
- Best QuAM is **Decision Tree Classifier.**

Future work for regression:

- Other regression methods.
- Polynomial features.

Future work in general:

- More data.
- Hyperparameters.
- Ensemble learning.


Thanks for Watching!