STUDIEREN. WISSEN. MACHEN.

HOCHSCHULE DER MEDIEN

MODELING & VALIDATION

MYROSLAVA SENKO

Agenda

- Introduction
- Feature Selection
- Algorithm Selection and Model Tuning
- Conclusion
- Questions

HOCHSCHULE DER MEDIEN

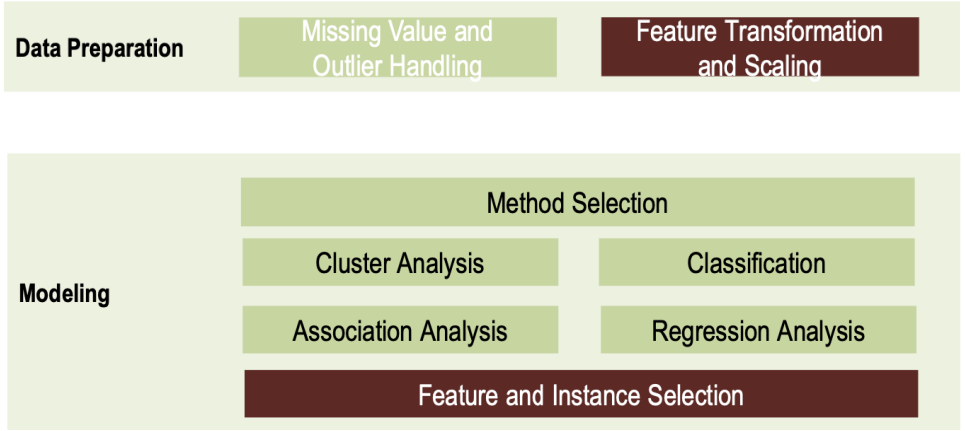
Preprocessing

Done:

› Removal of outliers and null values

To do:

› One Hot Encoding



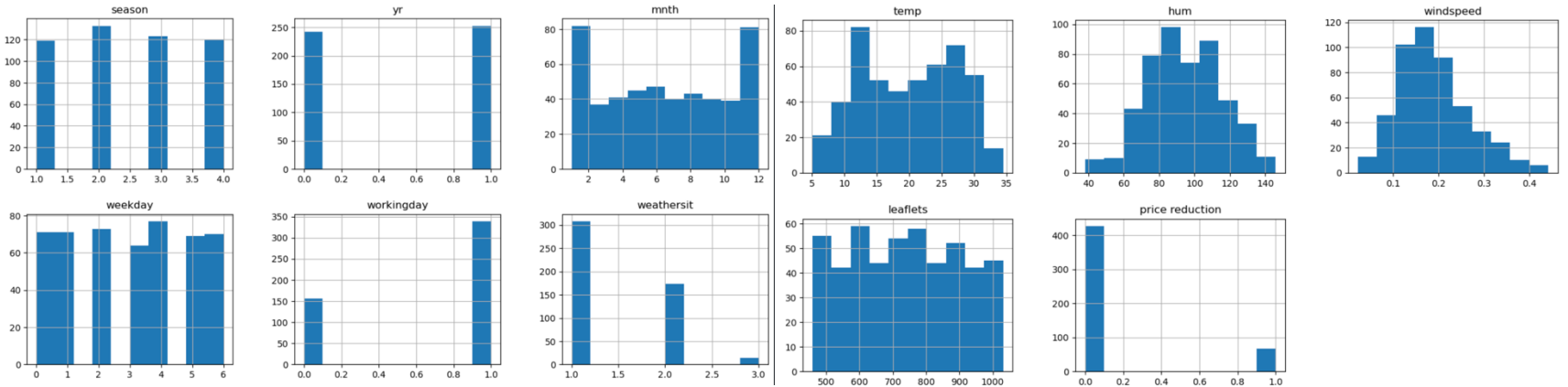
Modelling

› Removal of redundant features

› Algorithm Selection and Tuning

Train features

season	yr	mnth	weekday	workingday	weathersit	temp	hum	windspeed	leaflets	price reduction
2.0	0	6	5	1.0	1.0	24.8000	53.12505	0.253121	991.0	0.0
4.0	1	11	4	1.0	2.0	12.8667	93.06255	0.152987	601.0	0.0
1.0	1	1	2	1.0	1.0	6.0000	66.18750	0.365671	549.0	0.0
2.0	1	4	1	0.0	1.0	26.5667	84.25005	0.284829	740.0	0.0
1.0	1	3	6	0.0	2.0	20.5667	113.37495	0.110704	773.0	1.0
...	...	...	...	...	...	...	...	...	...	...
3.0	0	9	2	1.0	3.0	21.6000	133.04355	0.343943	763.0	0.0
2.0	0	4	5	1.0	2.0	13.4333	125.43750	0.226992	907.0	0.0
3.0	1	8	1	1.0	2.0	30.1000	98.12505	0.129354	861.0	1.0

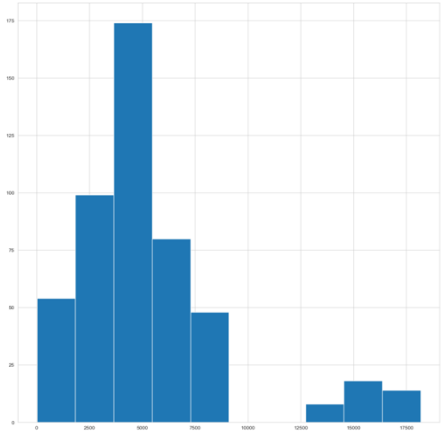


Deleted features

dteday	instant	casual	registered
2011-06-03	154	898.0	4414.0
2012-11-15	685	320.0	5125.0
2012-01-03	368	89.0	2147.0
2012-04-16	472	1198.0	5172.0
2012-03-17	442	3155.0	4681.0
...	...	...	...
2011-09-06	249	204.0	2506.0
2011-04-08	98	172.0	1299.0
2012-08-06	584	1233.0	5780.0

Label

cnt
5312.0
5445.0
2236.0
6370.0
7836.0
...
2710.0
1471.0
7013.0

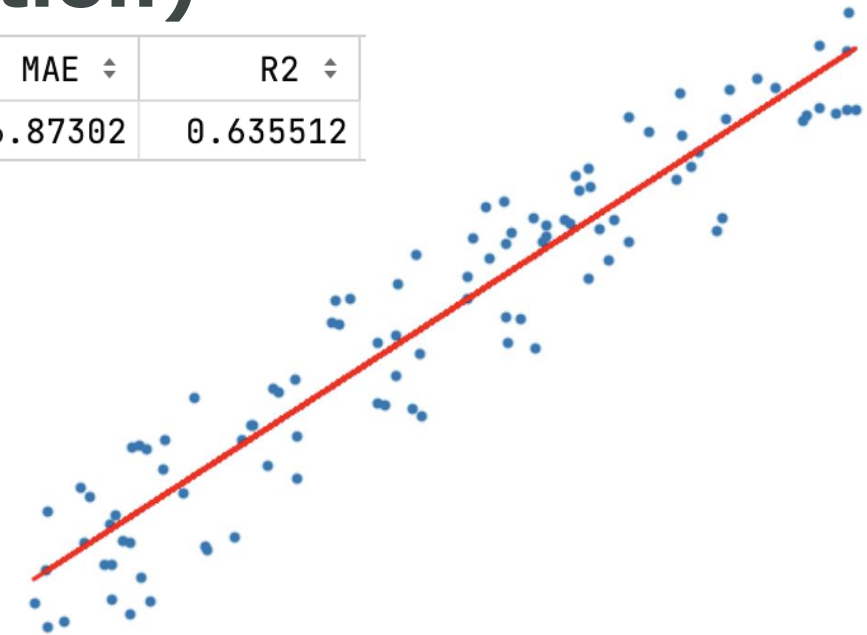


Deleting features which could distort the results.

Linear regression (preprocessed data without feature selection)

Model	MAE	R2
Linear regression (no selection)	1716.87302	0.635512

Linear regression without feature selection to set up a base model.



HOCHSCHULE
DER MEDIEN

Feature Selection

Nominal features

season	mnth	weekday	weathersit
2.0	6	5	1.0
4.0	11	4	2.0
1.0	1	2	1.0
2.0	4	1	1.0
1.0	3	6	2.0
3.0	9	3	1.0
1.0	2	5	1.0
4.0	10	4	2.0
4.0	10	6	2.0

season	mnth	weekday	weathersit
spring	June	Friday	Clear
autumn	November	Thursday	Mist
winter	January	Tuesday	Clear
spring	April	Monday	Clear
winter	March	Saturday	Mist
summer	September	Wednesday	Clear
winter	February	Friday	Clear
autumn	October	Thursday	Mist
autumn	October	Saturday	Mist

3	season	season (1:winter, 2:spring, 3:summer, 4:fall)
5	mnth	month (1 to 12)
7	weekday	day of the week
9	weathersit	weather situation: 1: Clear, Few clouds, Partly cloudy, Partly cloudy 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

Nominal features:

Feature variance

season	0.731313
mnth	0.905051
weekday	0.818396
weathersit	0.377778

All the features have variance > 0.05.

	season	mnth	weekday	weathersit
count	495	495	424	495
unique	4	12	6	3
top	spring	June	Thursday	Clear
freq	133	47	77	308

This table shows number of unique values for each nominal feature.

Nominal features: Variance and One-Hot-Encoding

HOCHSCHULE DER MEDIEN

Value variance

mnth_April	0.076122
mnth_August	0.079483
mnth_December	0.074428
mnth_February	0.076122
mnth_January	0.076122
mnth_July	0.074428
mnth_June	0.086108
mnth_March	0.069300
mnth_May	0.082812
mnth_November	0.076122
mnth_October	0.072727
mnth_September	0.074428
season_autumn	0.184026
season_spring	0.196892
season_summer	0.187118
season_winter	0.182980
weekday_Friday	0.120206
weekday_Monday	0.123110
weekday_Saturday	0.121662
weekday_Thursday	0.131624
weekday_Tuesday	0.125980
weekday_Wednesd…	0.112804
weathersit_Clear	0.235538
weathersit_Mist	0.227808
weathersit_Scat…	0.027539

One hot encoding

season_autumn	season_spring	season_summer	season_winter
0	1	0	0
1	0	0	0
0	0	0	1
0	0	1	0

Example of one-hot-encoding for the nominal value “season”. Before concatination one of columns will be deleted to avoid singularities.

› weathersit_Scattered has a variance of 0.027539, indicating that this weather situation occurs infrequently, compared to the others and thus doesn’t play a significant role for the model.

Nominal features: Chi-squared test

HOCHSCHULE DER MEDIEN

Feature

Feature	season	month	weekday	weathersit
season	(0)	≈0	0.99	0.127
mnth	-	(0)	0.99	0.012
weekday	-	-	(0)	0.325
weathersit	-	-	-	(0)

If the p-value -> 0, then the variables are dependent. In case of features it's a proof of collinearity, one of 2 features should be deleted.

- 0 - dependent
- 1 - independent

Feature	label
season	≈0
mnth	≈0 (< season)
weekday	0.97
weathersit	0.044

If the p-value -> 0, then the variables are dependent.
In case of label it's a proof of correlation, and is a sign that feature is important.

If the p-value -> 1, then the feature and label are independent.
The feature has no influence on the result and can be deleted.

Nominal features: Result features

HOCHSCHULE DER MEDIEN

mnth

mnth_August	mnth_December	mnth_February	mnth_January	mnth_July	mnth_June	mnth_March	mnth_May	mnth_November	mnth_October	mnth_September
0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0
0	0	0	1	0	0	0	0	0	0	0

weathersit

weathersit_Mist	weathersit_Scattered
0	0
1	0
0	0

Nominal features that will be used for modelling.

Numeric features: MinMaxScaler

	yr	workingday	temp	hum	windspeed	leaflets	price reduction
count	495.000000	495.000000	495.000000	495.000000	495.000000	495.000000	495.000000
mean	0.511111	0.684848	0.504746	0.521657	0.400560	0.485820	0.137374
std	0.500382	0.465046	0.248246	0.194587	0.186796	0.283377	0.344590
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.282633	0.381961	0.268526	0.243007	0.000000
50%	1.000000	1.000000	0.515322	0.514501	0.372412	0.480769	0.000000
75%	1.000000	1.000000	0.723609	0.660659	0.501922	0.718531	0.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Scaling all the features to perform Variance feature selection.
After scaling all the features are in the range [0,1].

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Numeric features: Variance

yr	0.250382
workingday	0.216268
temp	0.061626
hum	0.037864
windspeed	0.034893
leaflets	0.080303
price reduction	0.118742

Variance of features “hum” and “windspeed” is less than 0.05. If the features don’t perform well in correlation/collinearity testing, they will be deleted.

Numeric features

HOCHSCHULE DER MEDIEN

Correlation

temp	0.657667
yr	0.295124
workingday	0.001501
price reduction	-0.014675
leaflets	-0.031639
hum	-0.128074
windspeed	-0.206001

Let's take 5 features with the highest correlation to label. Features "workingday" and "price reduction" can be deleted.

Chosen features

temp	0.657667
yr	0.295124
windspeed	-0.206001
hum	-0.128074
leaflets	-0.031639

Features that will be used for modelling.

Numeric features

HOCHSCHULE DER MEDIEN

Collinearity

features	VIF Factor
temp	4.461709
yr	1.904048
windspeed	3.744500
hum	5.238580
leaflets	3.388039

- **VIF = 1**: not correlated to any other features
- **1 < VIF < 5**: moderately correlated
- **VIF >= 5**: a significant correlation with other features

	temp	yr	windspeed	hum	leaflets
temp	1.000000	0.027655	-0.180952	0.118373	0.002493
yr	0.027655	1.000000	-0.012301	-0.138729	-0.081810
windspeed	-0.180952	-0.012301	1.000000	-0.220864	-0.009016
hum	0.118373	-0.138729	-0.220864	1.000000	-0.044002
leaflets	0.002493	-0.081810	-0.009016	-0.044002	1.000000

Exploring collinearity to determine correlating features.

Results of feature selection

HOCHSCHULE DER MEDIEN

Chosen numeric features

- temp
- yr
- windspeed
- hum
- leaflets

Chosen nominal features

- mnth_August
- mnth_December
- mnth_February
- mnth_January
- mnth_July
- mnth_June
- mnth_March
- mnth_March
- mnth_May
- mnth_November
- mnth_October
- mnth_September
- weathersit_Mist
- weathersit_Scattered

We concatenate the following features and will use them for modelling.

HOCHSCHULE
DER MEDIEN

Algorithm Selection and Model Tuning

Algorithm Selection and Model Tuning

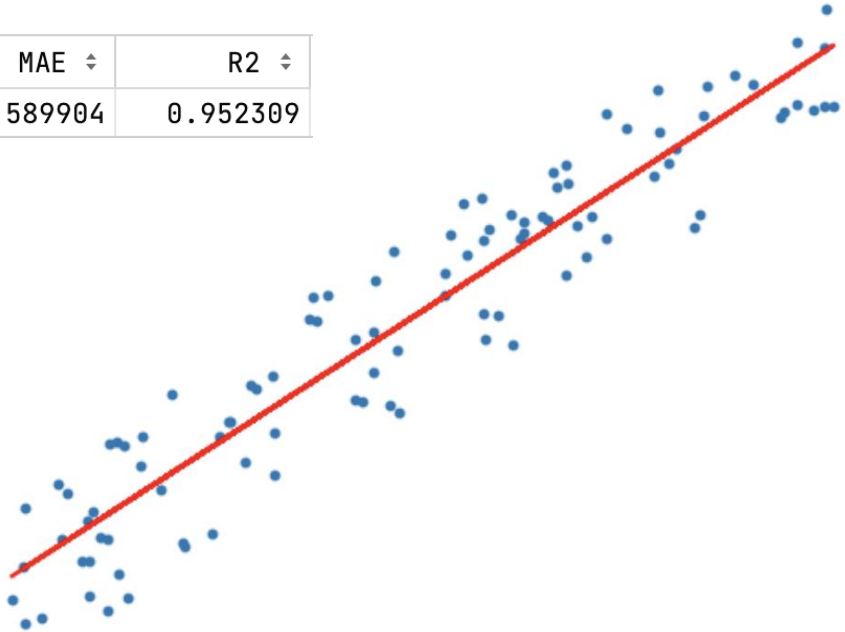
HOCHSCHULE DER MEDIEN

- Linear Regression
- Polynomial Regression
- K-nearest-Neighbors-Regression
- Decision Tree Regression

Linear Regression

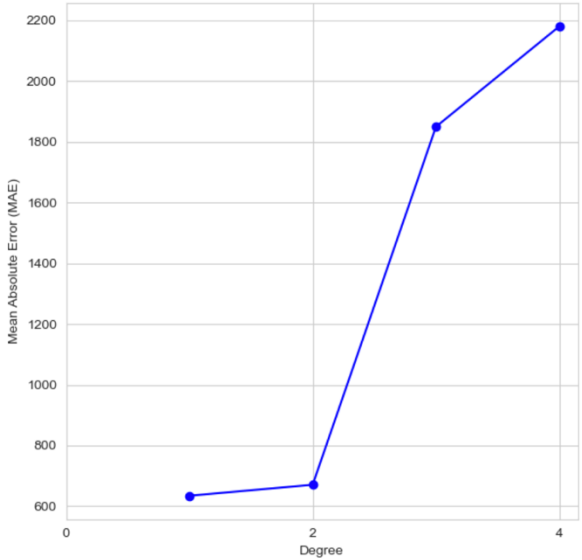
Model	MAE	R2
Linear regression (selected features)	634.589904	0.952309

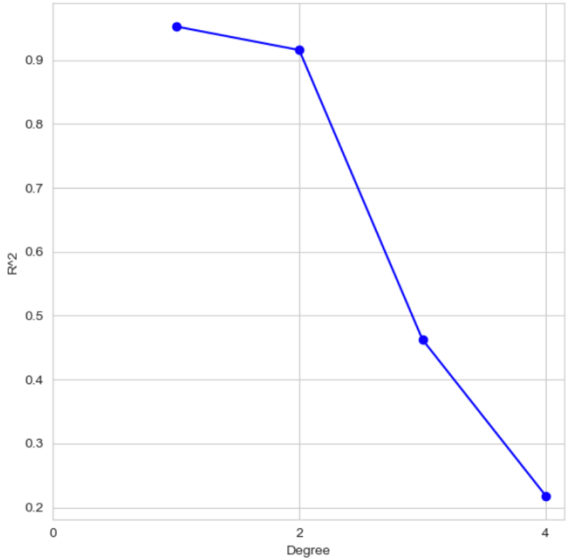
+32% more accuracy,
after selecting features.



Polynomial Regression

	Model	MAE	R2
0	Linear regression (selected features)	634.589904	0.952309
1	Polynomial regression (degree: 2)	670.648535	0.915439
2	Decision tree (Max Depth: 6)	780.020660	0.927206
3	KNN Regression (Number of Neighbors: 12)	665.016214	0.946123

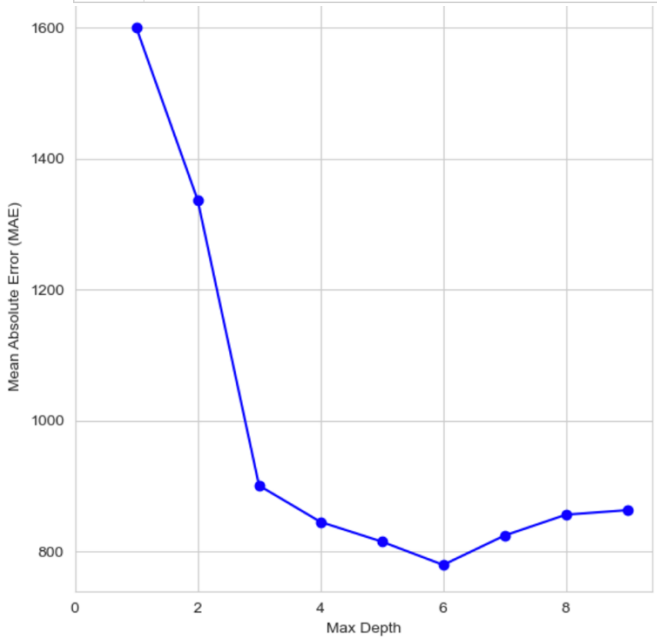


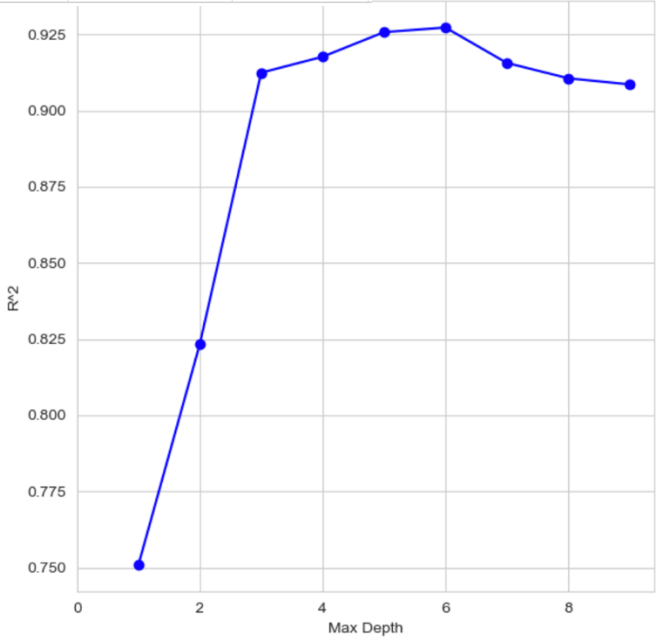


- Polynomial Regression with degree 2 has the best result.
- Degree 1 would be Linear regression and 3 has a bad performance.
- Decent MAE and R^2 values.

Decision Tree Regression

	Model	MAE	R2
0	Linear regression (selected features)	634.589904	0.952309
1	Polynomial regression (degree: 2)	670.648535	0.915439
2	Decision tree (Max Depth: 6)	780.020660	0.927206
3	KNN Regression (Number of Neighbors: 12)	665.016214	0.946123

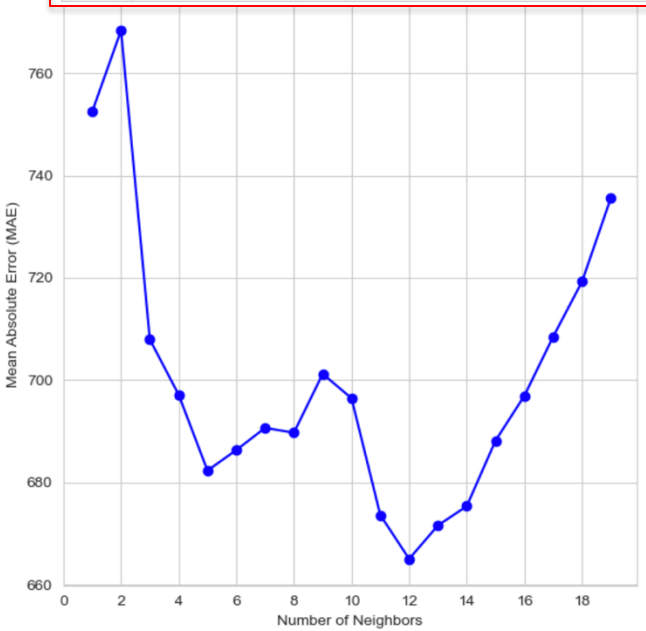


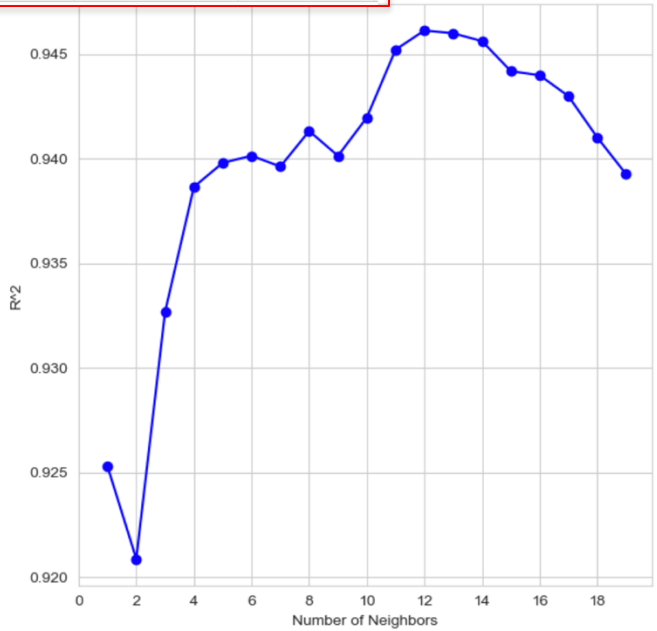


- Decision Tree with maximal Depth of 6 has the best result.
- Max Depth of 1 wasn't sufficent
- After Max depth of 6 performance declined
- Good MAE and R^2 Values
- Max depth refers to the depth of the roots the deeper the root, the more granular the result.

k-nearest-Neighbor-Regression

	Model	MAE	R2
0	Linear regression (selected features)	634.589904	0.952309
1	Polynomial regression (degree: 2)	670.648535	0.915439
2	Decision tree (Max Depth: 6)	780.020660	0.927206
3	KNN Regression (Number of Neighbors: 12)	665.016214	0.946123





- The KNN Regression model achieved the best results with a configuration of 12 neighbors.
- Model with 2 Neigbors performed worst
- KNN-Model scored the second highest perfomance among the evaluated models
- KNN-Model makes a prediction based on the majority of the k-nearest neighbors of a data point.

Results

No selection

	Model	MAE	R2
0	Linear regression (no selection)	1716.873020	0.635512
1	KNN regression (no selection, neighbors: 3)	2170.631976	0.258834
2	Polynomial regression (no selection, degree: 2)	2092.486436	0.453127
3	Decision Tree (no selection, max_depth: 5)	806.155972	0.826247

- As the results shows feature selection and tuning achieved better results than simply no feature selection.

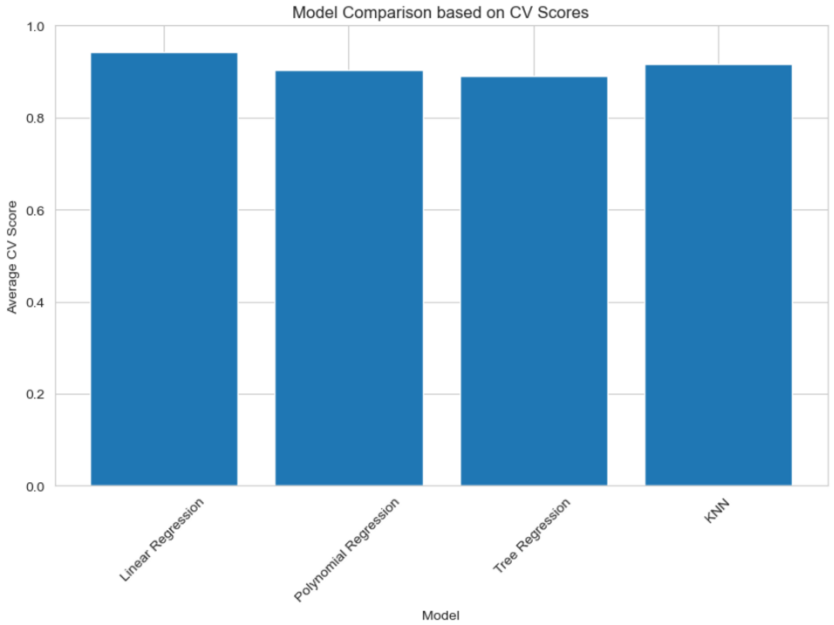
Feature selection + tuning

	Model	MAE	R2
0	Linear regression (selected features)	634.589904	0.952309
1	Polynomial regression (degree: 2)	670.648535	0.915439
2	Decision tree (Max Depth: 6)	780.020660	0.927206
3	KNN Regression (Number of Neighbors: 12)	665.016214	0.946123

- Feature Selection and tuning incresead the performance up to 32% accuracy.

Cross Validation

```
Cross Validation scores for Linear Regression:  [0.94676211 0.94230558 0.95018693 0.94229422 0.93439699]
Cross Validation scores for Polynomial Regression:  [0.9632467  0.9370062  0.95459113 0.92708458 0.7394573 ]
Cross Validation scores for Tree Regression:  [0.90401036 0.8711289  0.92139664 0.90895458 0.85123314]
Cross Validation scores for KNN:  [0.92645026 0.89568045 0.92421652 0.93490622 0.90336877]
```



Model	Average CV
Linear Regression	0.943189
Polynomial Regression	0.904277
Tree Regression	0.891345
KNN	0.916924

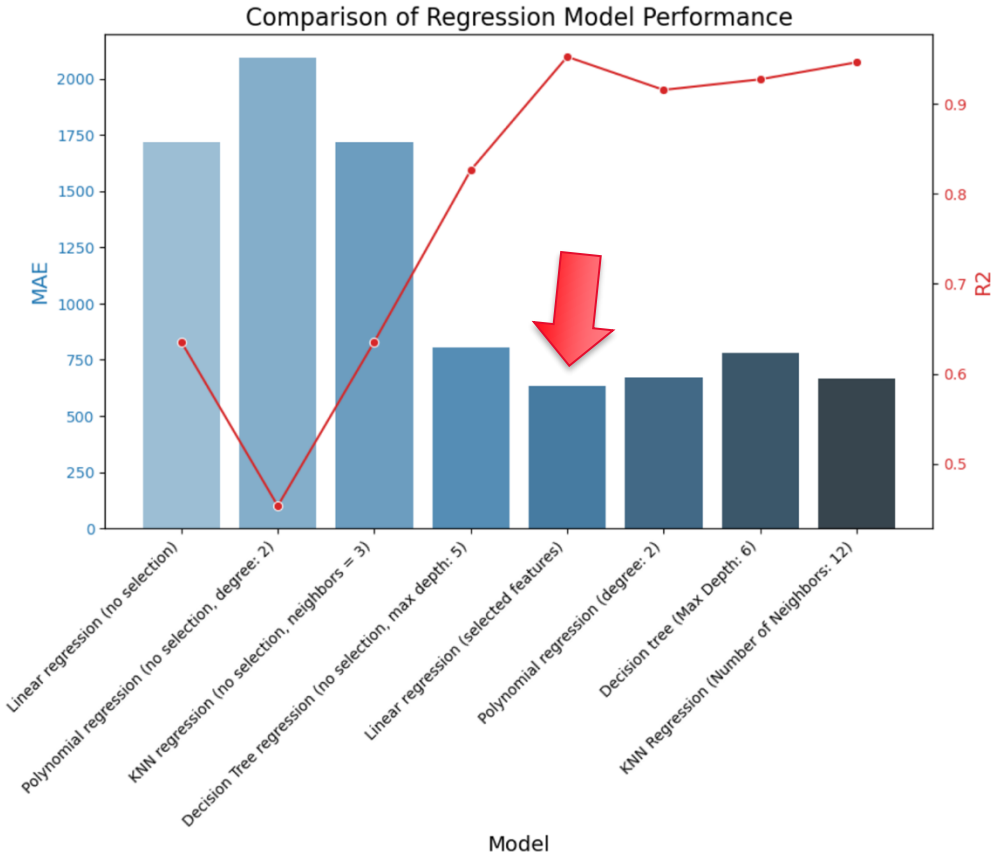
Cross Validation indicates that overfitting is not occurring.

Conclusion

Models with feature selction and tuning performed better than no feature selection Models.

Best perfoming Models:

1. Linear Regression with feature selection
2. KNN with Neigbors of 12
3. Polynomial Regression with degree 2
4. Decision Trees with max Depth 6
5. Decision Trees with no selection and max depth of 5



HOCHSCHULE
DER MEDIEN

Thanks for your attention!