Prediction of German House Prices

Knowledge Discovery (CENG-542)
Project Presentation

TOC

Objective
Preprocessing
Experiment-1
Experiment-2
Final Result

TOC

Objective
Preprocessing
Experiment-1
Experiment-2
Final Result

Objective

Objective: find patterns to determine prices for German houses based on different attributes

Problem: Regression of prices

Dataset: germany_housing_data_14.07.2020 (10552 rows, 24 columns)

Tools: Python + Libraries

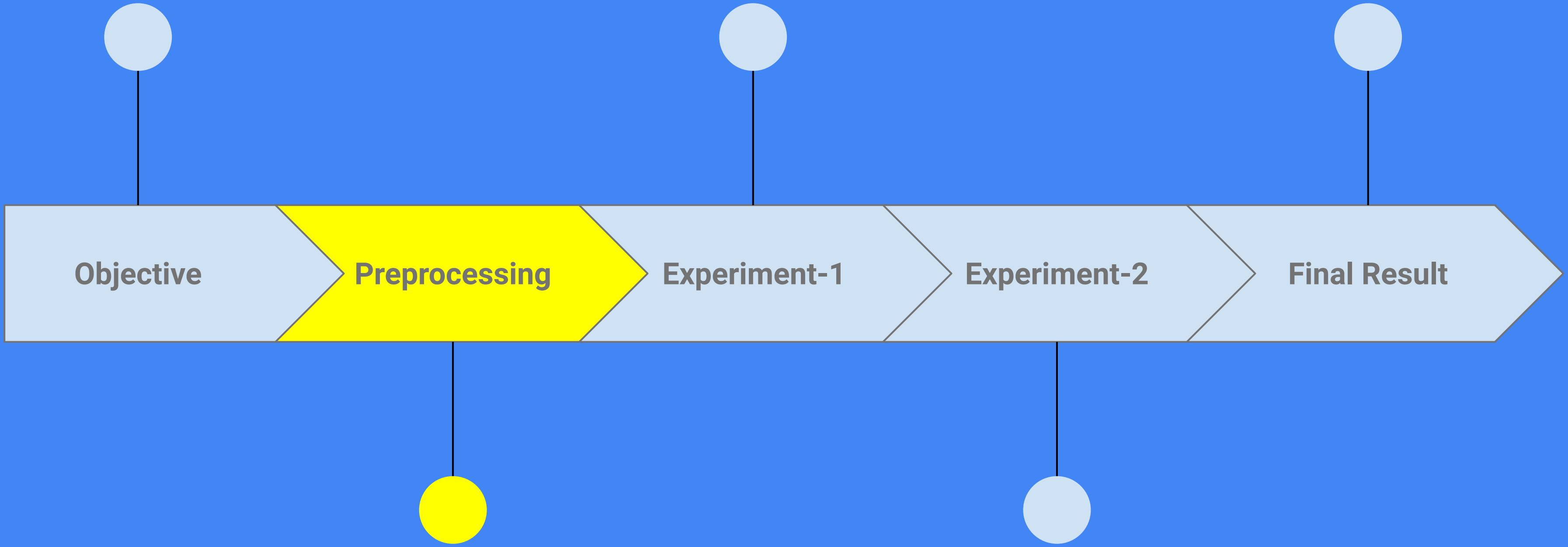
Getting to Know the Data

Dimension: 10552 x 25 (rows x columns)

Statistical Properties for Price: min = 0, mean = 556 685 , max = 13 000 000

3 Samples of dataset:

Type	Living_space	Lot	Usable_area	Bathrooms	Rooms	Year_built	State	…	Price
Mid-terrace house	140.0	890.0	NaN	3.0	4.0	1989.0	Sachsen	….	275000.0
Mid-terrace house	130.00	401.0	200.00	1	4	2014	Berlin	…	840000.0
Single dwelling	129.0	157.0.0	39.0	2.0	4.0	2018.0	Bayern	….	1075000.0


Objective
Preprocessing
Experiment-1
Experiment-2
Final Result

Preprocessing (a)

(1) Data Reduction (Dimensionality Reduction)

Drop features which have …

- … same information (State > City > Place)
- … > 33.333% NaN values (7 of 24 features)

(2) Data Cleaning (Missing Data)

Replace all remaining NaN values with **mode**

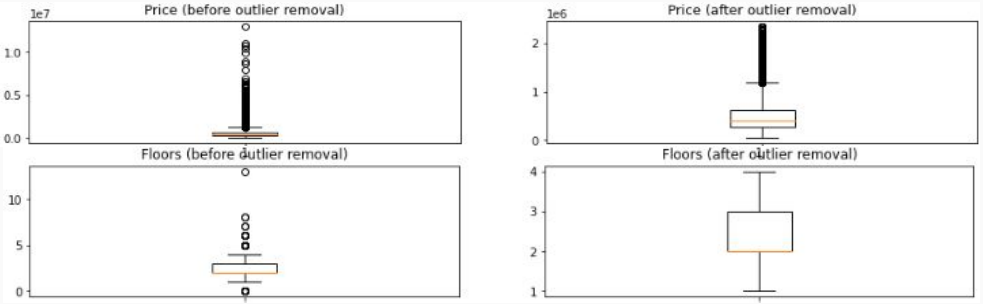
	nulls_amount	nulls_percentage
Energy_consumption	8119	76.94
Year_renovated	5203	49.31
Usable_area	4984	47.23
Energy_efficiency_class	4819	45.67
Bedrooms	3674	34.82
Free_of_Relation	3569	33.82
Energy_certificate_type	3526	33.42
Furnishing_quality	2726	25.83
Floors	2664	25.25
Garages	1960	18.57
Garagetype	1960	18.57
Bathrooms	1801	17.07
Energy_source	1227	11.63
Energy_certificate	755	7.16
Year_built	694	6.58
Heating	584	5.53
Type	402	3.81
Condition	323	3.06
Place	290	2.75
State	1	0.01
City	1	0.01
Price	0	0.00
Rooms	0	0.00
Lot	0	0.00
Living_space	0	0.00
Unnamed: 0	0	0.00

Figure 0 – NaN values in Dataset

Preprocessing (b)

(3) Data Cleaning (Outlier Removal)

removed 966 outliers using z-score (z > 3) and fixed price range of [50 000, 10 000 000]



	Before	After
Price (min.)	0 €	51 000 €
Price (mean)	556 685 €	499 877 €
Price (max.)	13 000 000 €	2 350 000 €

Figure 1 - Effect of Outlier Detection

Preprocessing (c)

(4) Data Transformation (One-hot encoding)

all categorical data got one-hot encoded

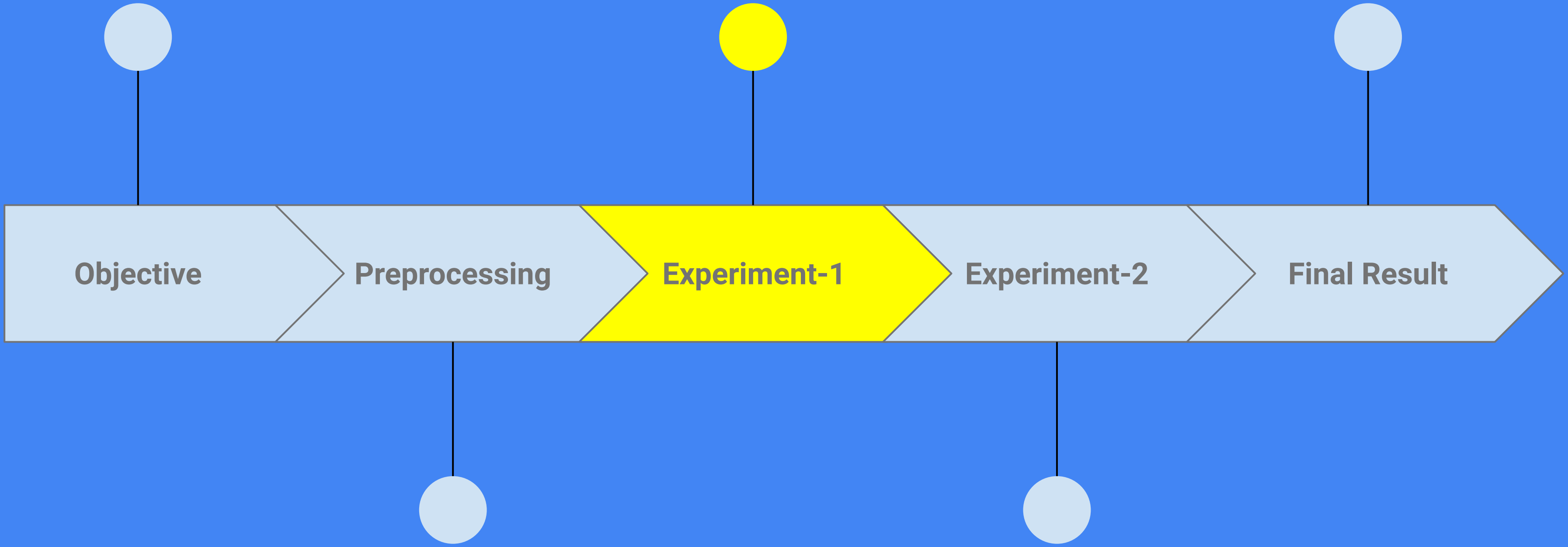
17 features → 170 features

(5) Data Reduction (Feature Selection)

Select only features with corr. factor >= 0.2

Feature	Living_space	Furnishing_quality_luxus	Bathrooms	Type_Villa	Rooms
Cor. factor	0.44	0.32	0.27	0.27	0.25

Figure 2 - Correlation Heatmap


Objective
Preprocessing
Experiment-1
Experiment-2
Final Result

Experiment-1

Procedure: linear regression models are trained, each on a different feature (80%, 20% data split)

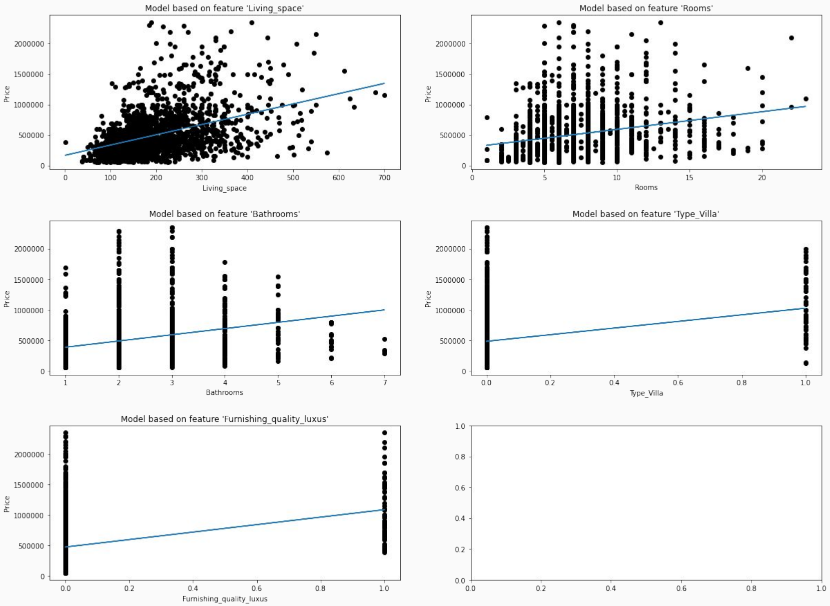


Figure 3 - Univariate Linear Regression Models

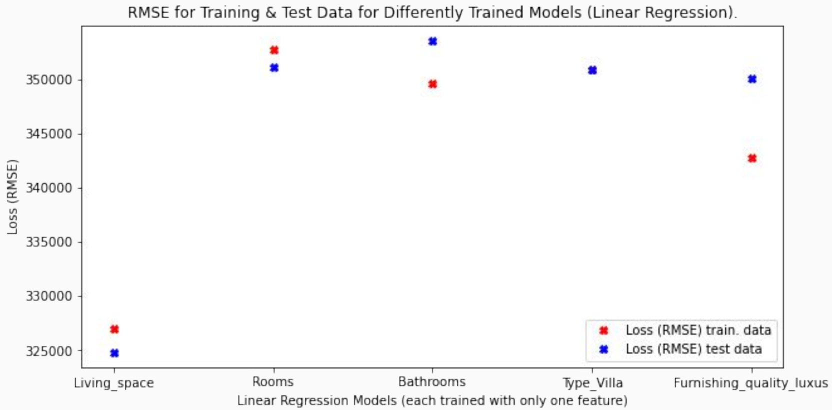
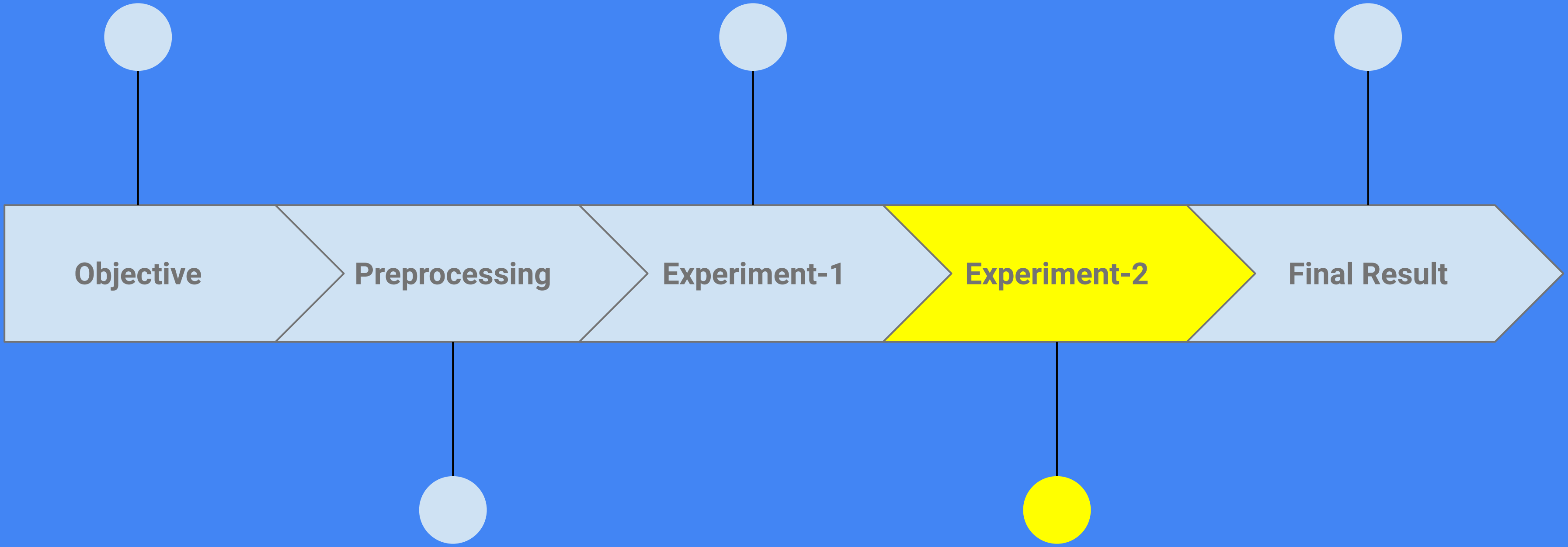


Figure 4 – Root-Mean-Squared-Errors


Objective
Preprocessing
Experiment-1
Experiment-2
Final Result

Experiment-2

Learning Algorithms:

(1) Lineare Regression
(2) K-nearest-neighbors (Regressor)
(3) Random Forest (Regressor)

Procedure:
For each learning algorithm try different amount of features: 1, 2, 3, 4, 5, or 170 (all)

→ 3 * 6 = 18 machine learning models

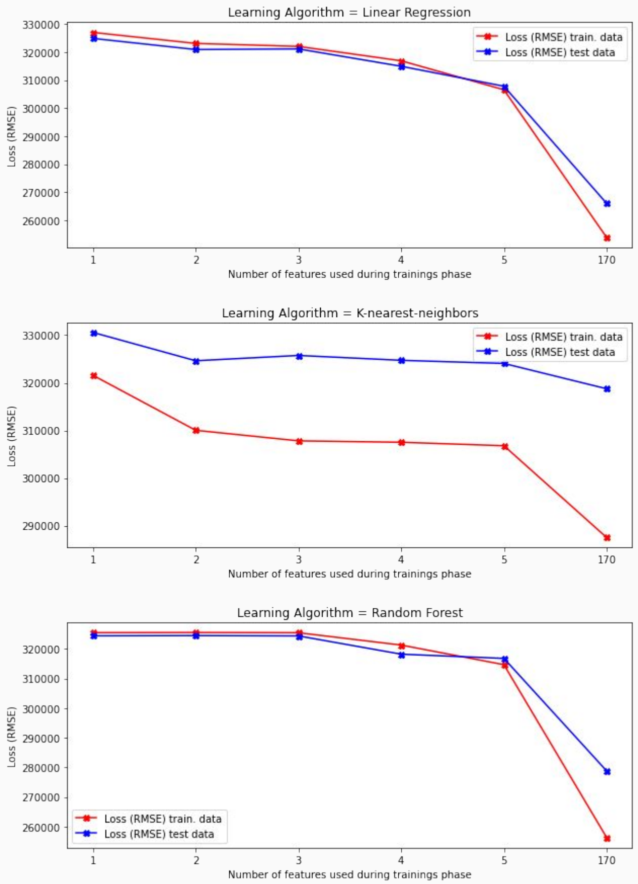
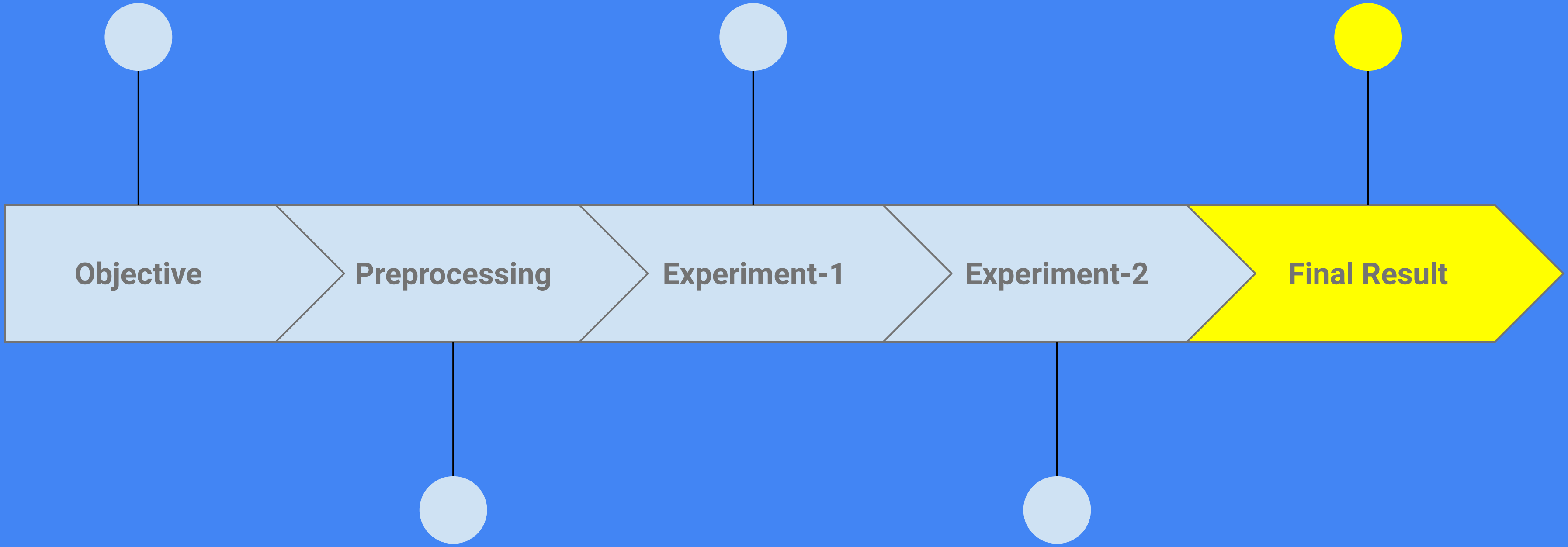


Figure 5 - Performance of Learning Algorithms


Objective
Preprocessing
Experiment-1
Experiment-2
Final Result

Final Result

Result: Best Model uses Multiple Lineare Regression (trained on all 170 features)
→ but: test data RMSE ~270000

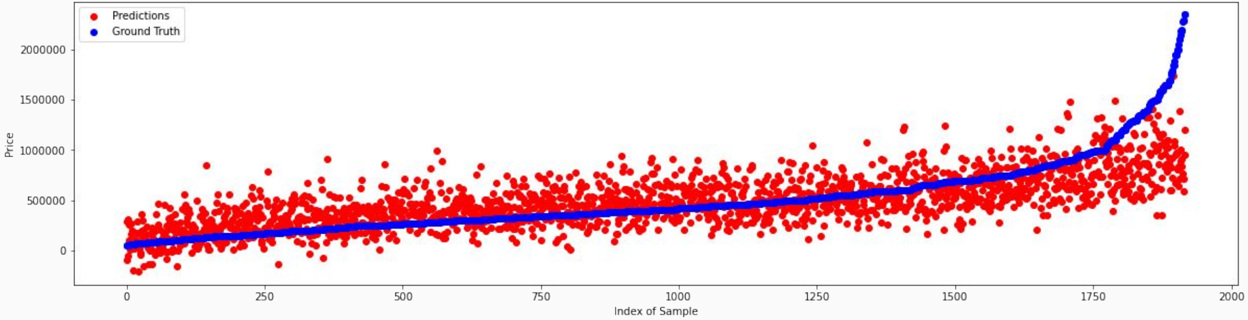


Figure 6 – Comparison between Ground Truth and Predictions

Conclusion: Even for the best model, the RMSE values are so high that it cannot make very accurate price predictions.

Thank you!

Questions?

About Me:

Joel Amarou Heuer
Freie Universität Berlin
Erasmus+ Student at IZTECH

joelheuer@iyte.edu.tr