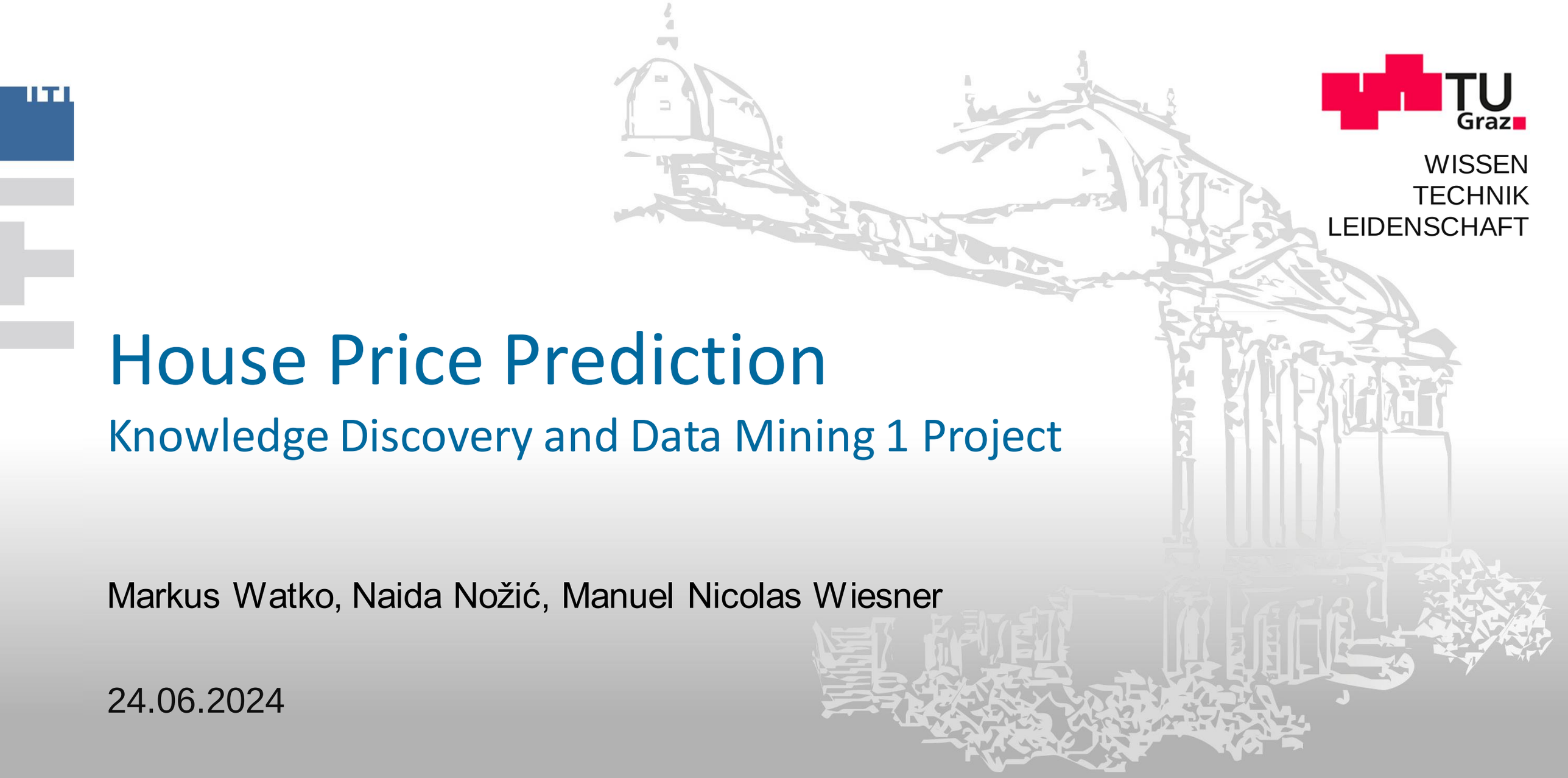


WISSEN
TECHNIK
LEIDENSCHAFT

House Price Prediction

Knowledge Discovery and Data Mining 1 Project

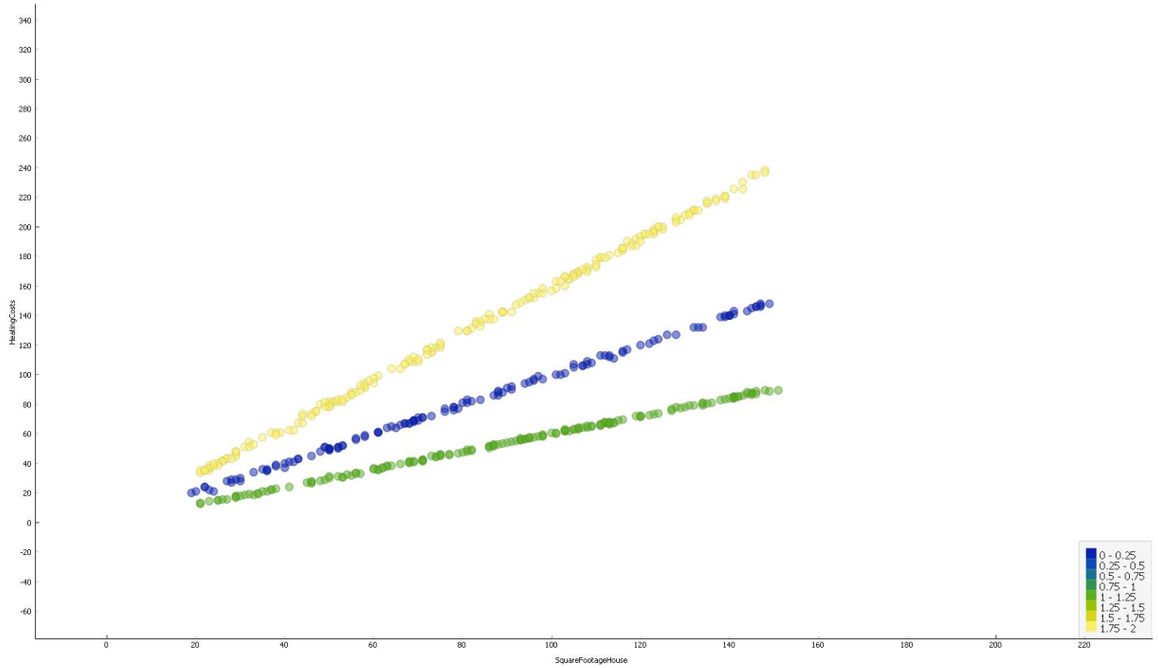
Markus Watko, Naida Nožić, Manuel Nicolas Wiesner

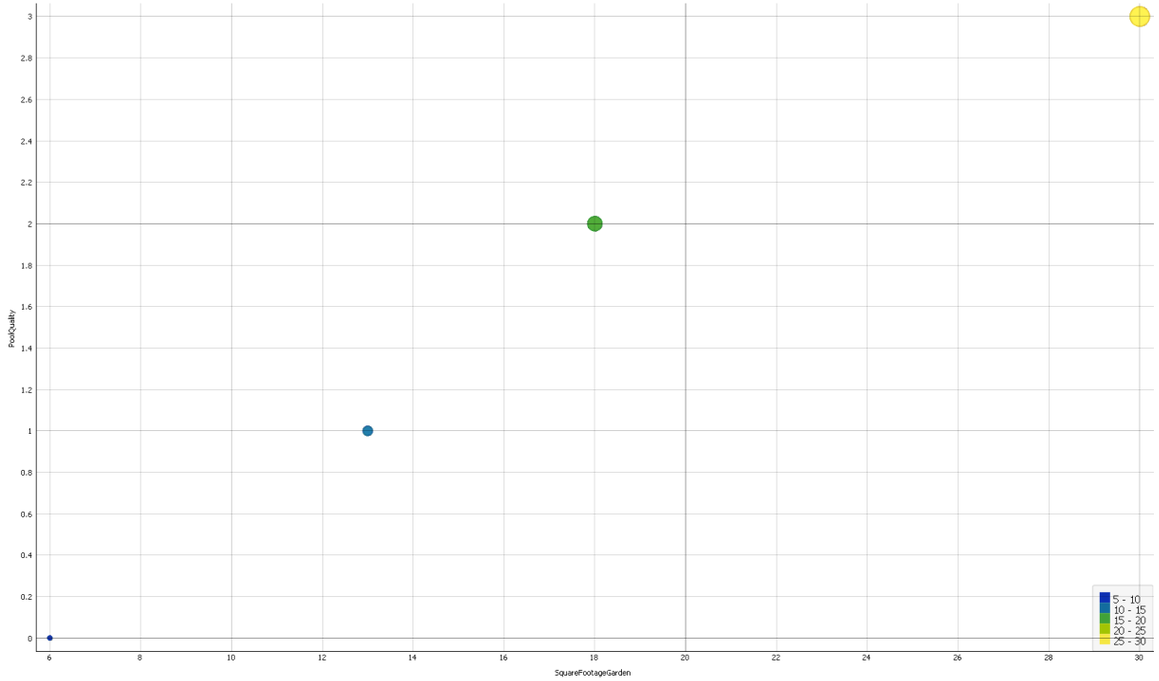
24.06.2024

Motivation

- Exploritory Data Analysis
- Data Preprocessing
- House Price Prediction
- Evaluation of the Machine Learning Model
- Question answers:
  1. If I have a budget of €100,000, what kind of houses will I be able to buy?
  2. How much money can I save if I decide not to get a house with a pool?

Exploring the data





Preprocessing

1. Bring data into shape
   - Drop completely empty rows or useless columns
   - Convert categorical values into enums
   - Convert numerical values to integers if possible
   - Simplify redundant values
   - Coerce some invalid values to null

2. Impute missing values
   - SFH => Bed-/Bathrooms
   - Price => Location
   - …

Machine Learning Models

- Linear Regression (LR)
  - baseline model to asses performance against
- Gradient Boosting Regression (GBR)
  - Best hyperparameters: 'learning_rate': 0.1, 'max_depth': 3, 'min_samples_split': 2, 'n_estimators': 200
- Random Forest (RF)
  - Best hyperparameters: 'max_depth': None, 'max_features': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 500
- Support Vector Regression (SVR)
  - best hyperparameters: 'C': 0.1, 'gamma': 0.1, 'kernel': 'linear

Training and Evaluation

- 80/20 dataset split to train and evaluate all models.
- 5 fold cross validation
- Performance evaluation metrics: MSE, MEA, R2 score, RMSE
- Results:
- **LR**: MSE: 2393, MAE: 39, RMSE: 48, R2 score: 0.77
- **SVR**: MSE: 3966, MAE: 49, RMSE: 63, R2 score : 0.61
- **RF**: MSE: 1052, RMSE: 34, MAE: 25, R2 score: 0.87
- **GBR**: MSE: 856, MAE: 20, RMSE: 31, R2 score 0.90

Discussion

Savings

Age <-> DateSinceForSale

Typos

Prices

HeatingType <-> HasPhotovoltaic <-> HeatingCost

Previous Owner

Windows

Houses < 100k