# Vendor Performance Benchmarking for ImmoScout24

Customer research at ImmoScout24 has shown vendors assess their performance on the market, in comparison with other vendors with similar real estate properties. Similarity of products is defined with a set of fixed rules. The comparison result is categorized as below, average or above average for the vendor. This thesis improves the current solution.

## Introduction

Performance is measured using leads. In our context leads are a summary of impressions tracked on the ImmoScout24 platform. Vendor performance is ranked in comparison with similar objects, based on number of leads. A, heuristically created, rule based approach is used to extract a predefined number of similar objects within a time frame, e.g. 3 similar objects within a month. Due to missing values 56% of target vendors have not sufficient comparable properties. The search algorithm is a naive based K-nearest algorithm, which is time consuming with big data.
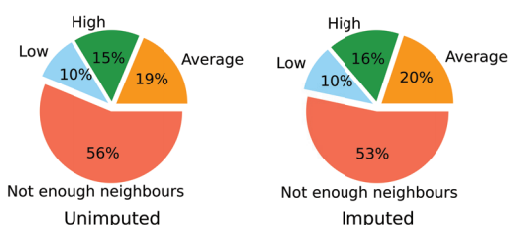
## Goals

- Improving the data quality
- Discovering clusters of property features
- Dependency and correlation analysis
- Efficiency of the search algorithm

## Results

### Improving the data quality

Similar objects are defined using living space, price and number of rooms. The data has 18% missing values on living space, 7% on price and 6% on number of rooms. Missing values diminish the number of similar properties for comparison. In this thesis we explored heuristic and regression methods for imputation. This yields decreasing the total percentage of missing values to 1% for living space, 1% for price and 1% for number of rooms. Imputing missing values diminished the amount of properties with not enough neighbors within a month time span to 53%.

### Dependency and correlation analysis

In the current setting, the performance of a vendor is related to object features. The assumption is the total number of leads generated by a real estate advert is related to living space, price and number of rooms. Customer research at ImmoScout24 has shown that vendors heuristically compare their properties based on these features. We have analyzed this relationship statistically. We show that based on the Hilbert-Schmidt Independence Criterion the output is dependent on the input variables. Moreover, Pearson correlation analysis shows linear correlation between input features.

### Efficiency of the search algorithm

The current search algorithm finds similar properties by applying a kNN search on clustered data. Clusters are based on offer type, property category and the object ZIP code. To improve the search algorithm we apply FAISS, which is an open-source library for efficient similarity search developed by Facebook. We ran the algorithms on a MacBook with a 2.6 GHz 6-Core Intel Core i7 and 32 GB of RAM. Comparing to kNN, FAISS extracts the nearest neighbors very fast. Comparing target properties within a 30 days time span, the execution time for the base search algorithm versus FAISS is 86 and .5 seconds respectively.

Because we explored different imputation methods, working on improving the data quality took longer than planned. We therefore decided to not pursue the goal ‚discover cluster of property features'. This was decided in accordance with our risk assessment. The other goals were all achieved.

Frédéric Nicolas Lehmann



Leads classification (using data from a 30 days span)
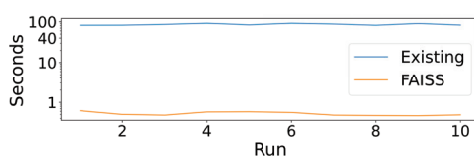


Run time comparison (using data from a 30 days span)