Coursera - IBM Data Science Specialization

Sebastian Mack

mack.seb@gmail.com

Abstract

dasd

With

Keyword: Data Science; Prediction; GBM; Tuning

# Introduction

The background for this capstone project is to have the opportunity to be as creative as possible and come up with ideas to leverage the Foursquare location data to come up with a problem that can make a benefit from the Foursquare location data.

* 1. **Project Overview**

As the field of study, I selected an interesting topic from the rental industry that caught my attention when I was exploring topics on the data science platform Kaggle. The initiator of this contest is RentHop (a portfolio company of TwoSigma Ventures) which has the objective to make apartment search smarter by using data to sort rental listings by quality (refer to <https://www.renthop.com/>).

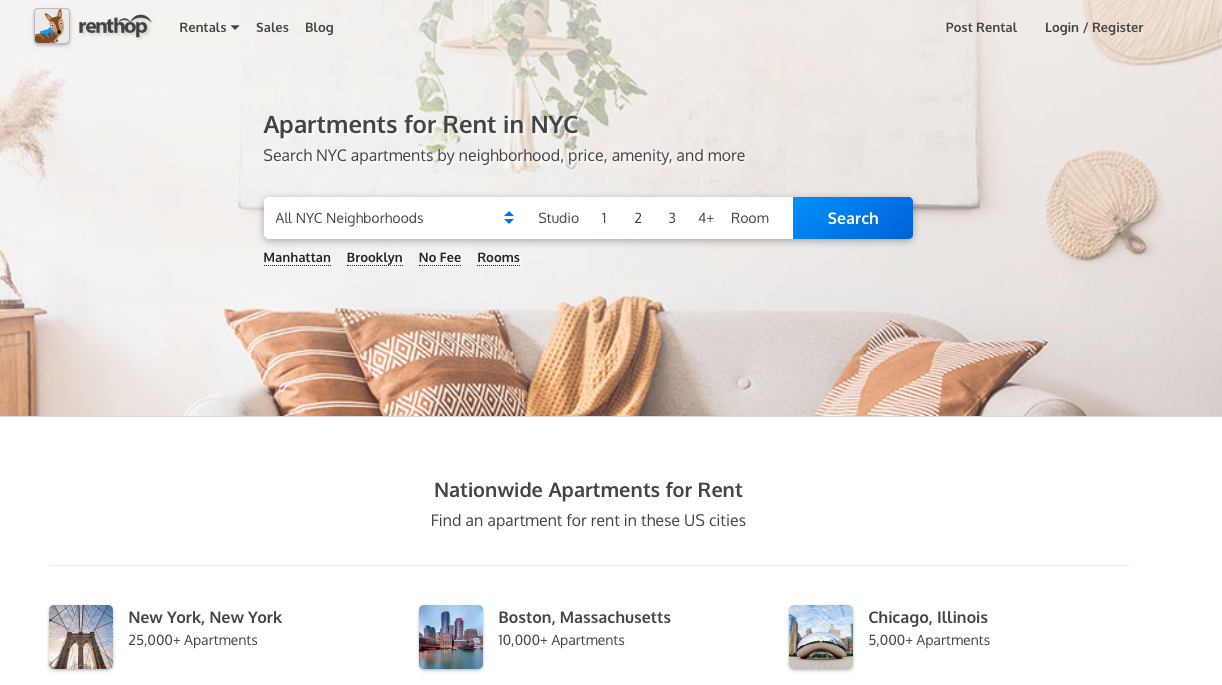
****

Fig. 1: Renthop Portal

* 1. **Problem Statement**

In this project we want to predict how popular an apartment rental listing is based on the listing content like text description, photos, number of bedrooms, price, etc. The data comes from renthop.com, an apartment listing website. These apartments are located in New York City. The target variable, interest level, is defined by the number of inquiries a listing has in the duration that the listing was live on the site. For me personally this problem is very interesting because it represents a real life problem that not only Renthop but many other companies are facing right now. In this situation companies are already acquiring and collecting data with their existing services but still struggle to create a business value out of it. In this particular case we can see that by developing a model that can predict how much interest a new rental listing on RentHop will receive, new business values can be proposed. Both the consumer and the merchants could benefit from such a situation. Furthermore, the offering company could gain new clients with this value adding service.

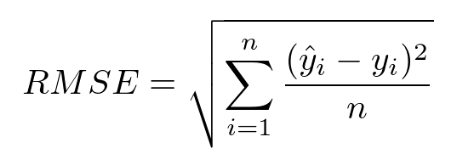
* 1. **Approach**

The problem that is to be solved can be described as a supervised machine learning problem because the model will be trained based on a given feature. Since this target variable has categorical values it can be further characterized as a classification problem. A problem of this type can be solved and modeled with various approaches but in this study the most promising will be applied which will be a desicion tree model. I will outline the most important steps within my theoretical workflow in order to find a solution for the described problem. A structured approach will be helpful for a reasonable result and to have a scientific discussion on the final model. My planned workflow includes the following steps:

* + 1. **Exploratory Data Analysis (EDA):** As a first step I will explore the provided data and make an analysis. This includes summarizing properties and visualizing important outcomes. It will be also very useful to identify features that are relevant for the model and also to give hints for transformations that are required for fitting the data.
    2. **Feature Engineering**: Within this step the knowledge gained from the previous step will be applied to clean the data set and to select important features as well as to define new ones.
    3. **Train Baseline Model**: When the previous step is completed, the obtained transformed and extended dataset can be used to train the baseline model (or benchmarking model). It will be an implementation of an Ensemble model with gradient boosting for classification.
    4. **Tune parameters**: Since there are lots of parameters available in order to train a sophisticated model, it will be necessary to repeat some of the steps and fine tune the model until it is able to produce the desired scores.
    5. **Evaluate Metrics**: In the last step, the results and scores of all generated models will be evaluated and compared to one another.

* 1. **Metrics**

The chosen evaluation metric will be the root mean squared error (RMSE)



1. **Data Description**

For this project the publicly available data for the” Two Sigma Connect: Rental Listing Inquiries” Kaggle competition will be considered. It consists out of the following files:

(<https://www.kaggle.com/c/two-sigma-connect-rental-listing-inquiries/data>)

* train.json: the training set
* test.json: the test set
* samplesubmission.csv: a sample submission file

For both the train and the test set, the following features are provided in thecompetition data. In total we have 49352 datapoints.

* bathrooms: number of bathrooms
* bedrooms: number of bathrooms
* buildingid
* created
* description
* displayaddress
* features: a list of features about this apartment
* latitude
* listingid
* longitude
* managerid
* photos: a list of photo links
* price: in USD
* streetaddress
* interestlevel: this is the target variable. It has 3 categories: ‘high’, ‘medium’, ‘low’

In the following figures we can find the head of the data frame as well as a statistical summary of the numerical features in the dataset.

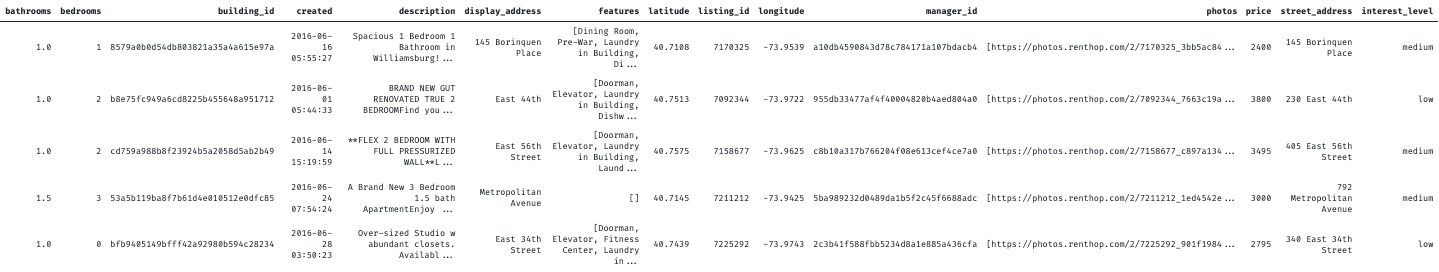


Fig. 2: Head of Competition Data

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

Fig. 3: Description Numeric Features Competition Data

In addition to the previously described data, we use the Foursquare api (<https://developer.foursquare.com/places-api>) to gather additional information for the respective geographic coordinates (latitude, longitude) that can be found for each row in the main data frames.

The hypothesis of this project is that we can use the foursquare api for exploring a location for improving the model. In order to do so, we have to specify an url and pass the coordinates of the relevant data point and make a request. The resulting json response can be extracted and put into a new pandas data frame. The new created features are:

* categories: number of unique foursquare venue categories
* distance: mean distance of foursquare venues

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

Fig. 4: Head of Crawled Foursquare Data

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

Fig. 5: Description of numeric features of Crawled Foursquare Data

Both of the two previously introduced data frames are merged into a single data frame for consistency. Afterwards additional steps like data cleaning and feature selection can be performed as well as engineering of features. We can extract with the help from pandas datetime module new features from the date and time. In addition, we can use word count techniques from the text feature. The total number of extracted additional features is only 6985 because of quota limits for free usage of the foursquare api.

1. **Methodology**

In this chapter we discuss the methodology that has been applied to the given problem. It includes the steps for preprocessing the data in order to address any abnormalities or characteristics. Furthermore, it documents the implemented metrics, algorithms and techniques. The general methodology is described by the Team Data Science Process ([TDSP](https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle)) shown in the following figure:

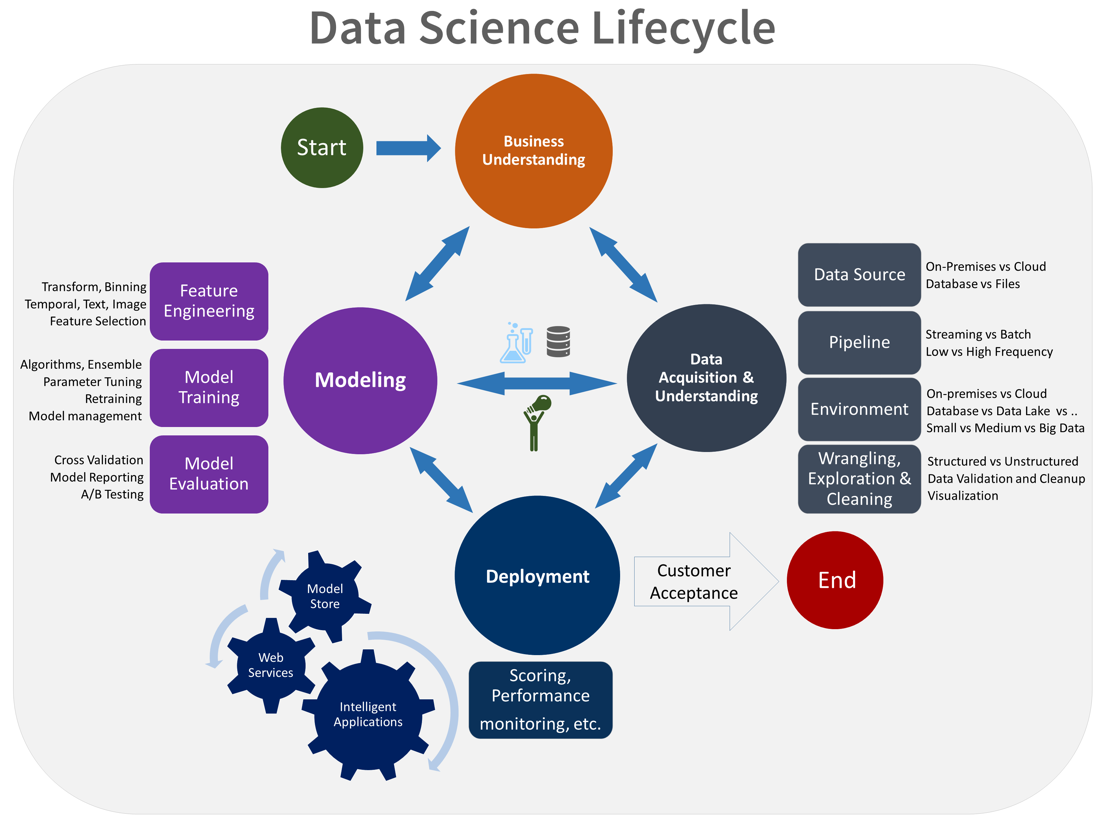


Fig. 6: Data Science Lifecycle ([TDSP](https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle))

* 1. **Exploratory Data Analysis (EDA)**

Before we begin to model the given problem, an analysis of the given data sets needs to be performed in order to understand which algorithms are going to be used in the next steps of the project. This section includes a data exploration which describes characteristic properties of the data as well as visualizations that help to summarize the most important outcomes.

As can be seen from the following figures, there are significantly more samples with low interest levels (5000) than medium (1500) and low (500) interest levels. The map plotted with folium illustrates this observation quite well.

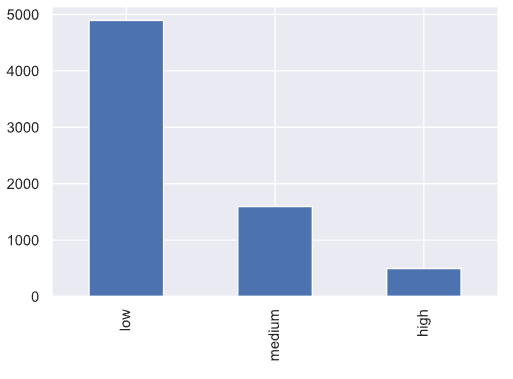
****

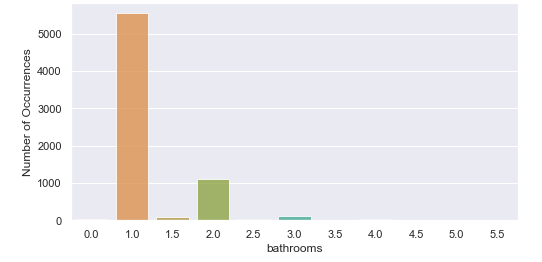
Fig. 7: Number of Occurrences of target

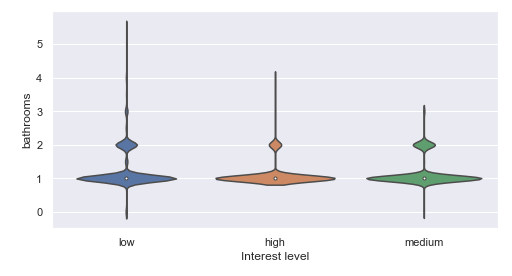
Ein Bild, das Karte enthält.

Automatisch generierte Beschreibung

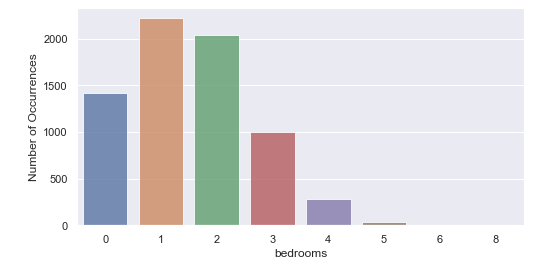
In order to get a better understanding for the predicting features we plot their distributions and relations to the target variable “interest level”.

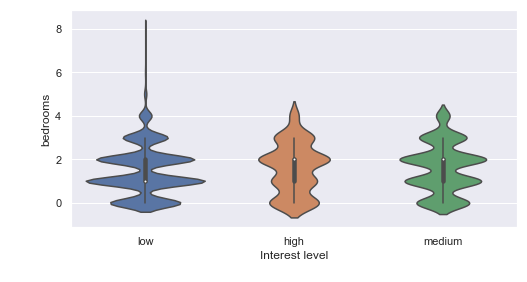
**Bathrooms:**





**Bedrooms:**

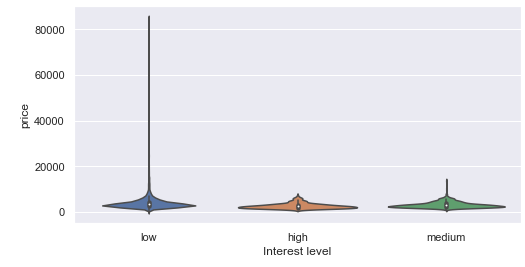




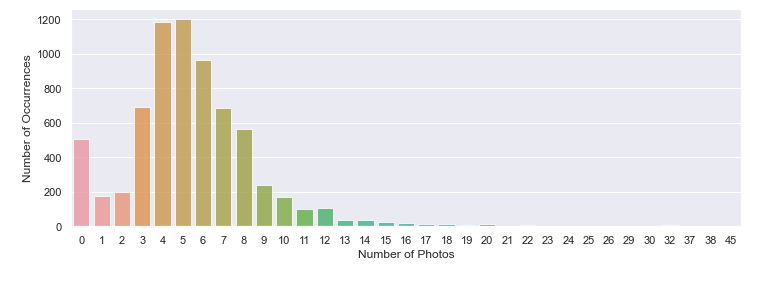
**Price:**

Ein Bild, das Tisch enthält.

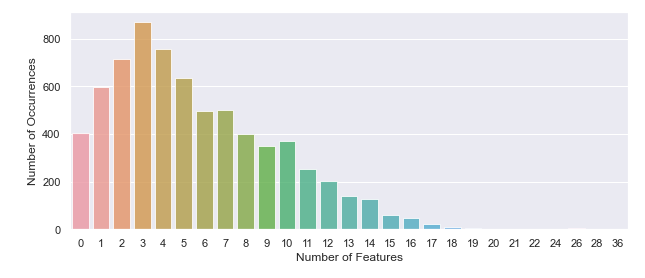
Automatisch generierte Beschreibung



**Number of photos:**



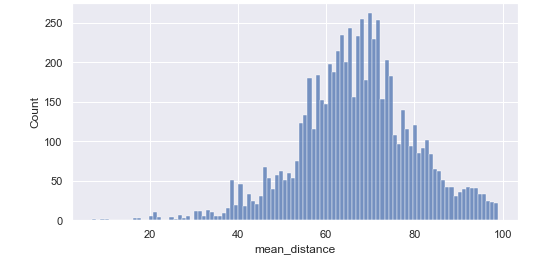
**Number of features:**

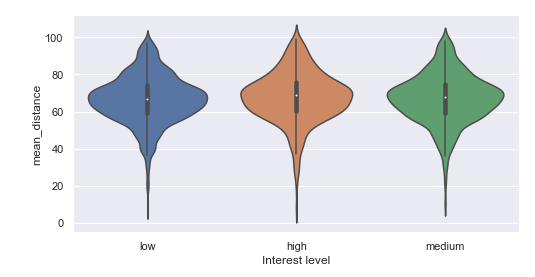


In addition, we examine the foursquare data features. For the geographical visualisations, green points represent a low value, yellow a medium value and red a high value.

**Mean distance of foursquare venues** Ein Bild, das Karte enthält.

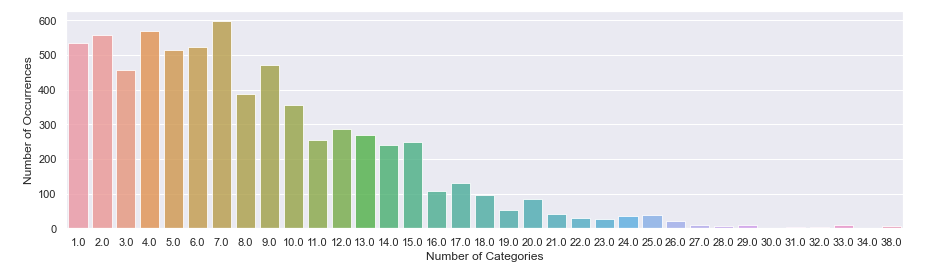
Automatisch generierte Beschreibung

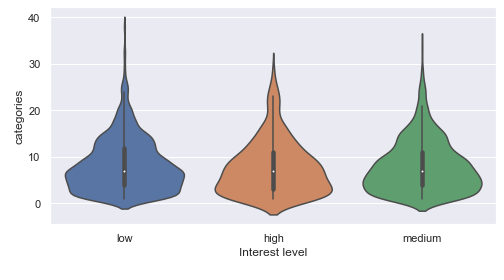




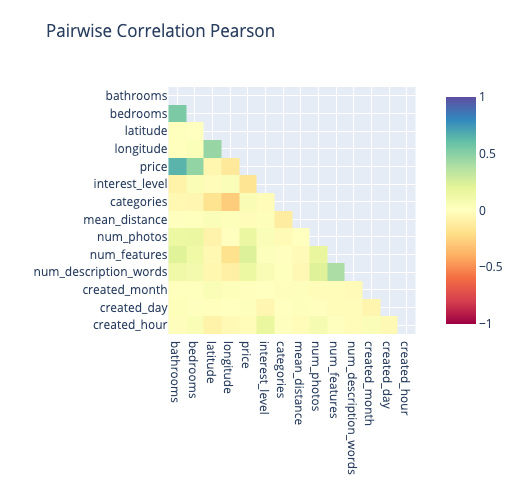
**Number of unique foursquare venue categories** Ein Bild, das Karte enthält.

Automatisch generierte Beschreibung





* 1. **Correlation**

****

Unfortunately, the newly created features "categories" and "mean\_distance" have very low correlation values with the target variable (-0.03 and 0.02 respectively). Maybe the model can identify more complex patterns.

* 1. **Algorithms**

For the modelling process of this project, we will use the h2o.ai framework. H2O is a Java-based software for data modeling and general computing. The H2O software is many things, but the primary purpose of H2O is as a distributed (many machines), parallel (many CPUs), in memory (several hundred GBs Xmx) processing engine. (<http://docs.h2o.ai/h2o/latest-stable/h2o-py/docs/intro.html>)

In particular the H2OGradientBoostingEstimator will be used to train a baseline model and compare it to an advanced model with the newly created features. [Gradient boosting](https://en.wikipedia.org/wiki/Gradient_boosting) is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient boosted trees, which usually outperforms random forest. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

1. **Results**
2. **Discussion**
3. **Conclusion**

The Purpose of this project was to identify new features for rental listings that can be used to predict the interest level of a customer. For this reason two features have been proposed.