Coursera - IBM Data Science Specialization Capstone Project

Sebastian Mack

[1] Sebastian Mack
[2] mack.seb@gmail.com

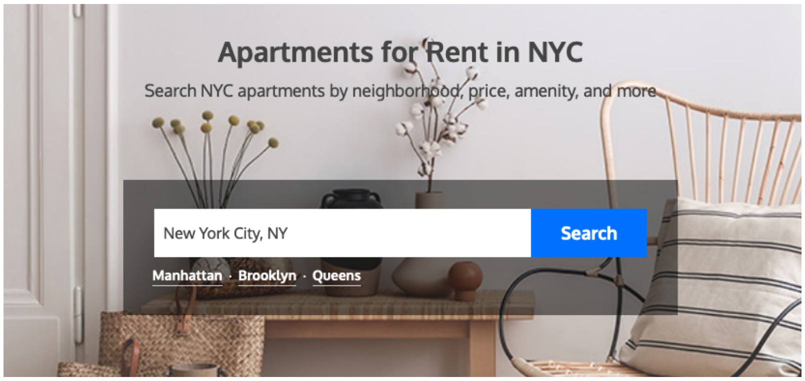
Abstract. Keywords: Data Science · Prediction · GBM · Tuning.

1 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 use the Foursquare location data to solve.

1.1 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 Two Sigma 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/).



Apartments for Rent in NYC

Search NYC apartments by neighborhood, price, amenity, and more

New York City, NY

Fig. 1. Renthop Platform

1.2 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.

1.3 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 target 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 theorectical 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 paramters**: Since there are lots of parameters available in order to train a suffisticated 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.

1.4 Metrics

The choosen evaluation metric will be the multi-class logarithmic loss as suggested by the competition rules.

2 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
- **sample_submission.csv** - a sample submission file

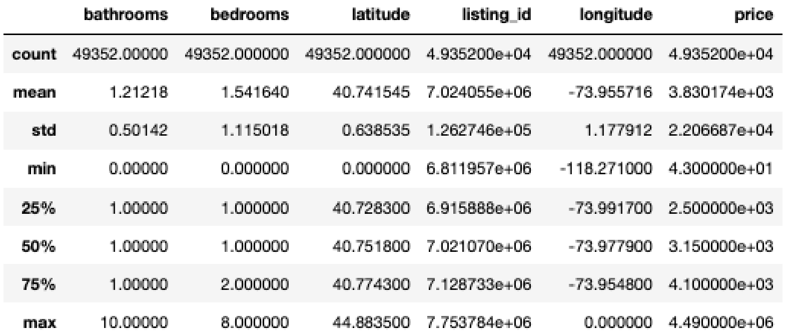
For both the train and the test set, the following features are provided in the competition data. In total we have 49352 datapoints.

- bathrooms: number of bathrooms
- bedrooms: number of bathrooms
- building_id
- created
- description
- display_address
- features: a list of features about this apartment
- latitude
- listing_id
- longitude
- manager_id
- photos: a list of photo links
- price: in USD
- street_address
- interest_level: this is the target variable. It has 3 categories: 'high', 'medium', 'low'

In figure 2 we can see a statistical summary of the numerical features in the dataset. 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 dataframe. The new created features are:

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



	bathrooms	bedrooms	latitude	listing_id	longitude	price
count	49352.00000	49352.000000	49352.000000	4.935200e+04	49352.000000	4.935200e+04
mean	1.21218	1.541640	40.741545	7.024055e+06	-73.955716	3.830174e+03
std	0.50142	1.115018	0.638535	1.262746e+05	1.177912	2.206687e+04
min	0.00000	0.000000	0.000000	6.811957e+06	-118.271000	4.300000e+01
25%	1.00000	1.000000	40.728300	6.915888e+06	-73.991700	2.500000e+03
50%	1.00000	1.000000	40.751800	7.021070e+06	-73.977900	3.150000e+03
75%	1.00000	2.000000	40.774300	7.128733e+06	-73.954800	4.100000e+03
max	10.00000	8.000000	44.883500	7.753784e+06	0.000000	4.490000e+06

Fig. 2. Summary of numerical features

The two dataframes are merged into a single dataframe 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 modul new features from the date and time. In addition we can use word count techniques from the text feature.

3 Methodology

In this chapter we discuss the methodology that has been applied to the given problem. It includes the steps for prepocessing the data in order to address any abnormalities or characteristics. Furthermore, it documents the implemented metrics, algorithms and techniques.

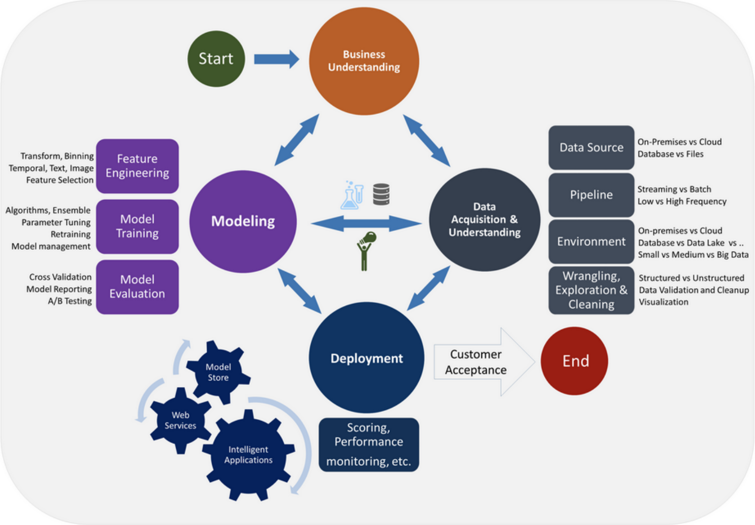


Fig. 3. Data Science Lifecycle

3.1 Exploratory Data Analysis

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 (35000) than medium (10000) and low (4000) interest levels. The map plotted with folium illustrates this observation quite well.

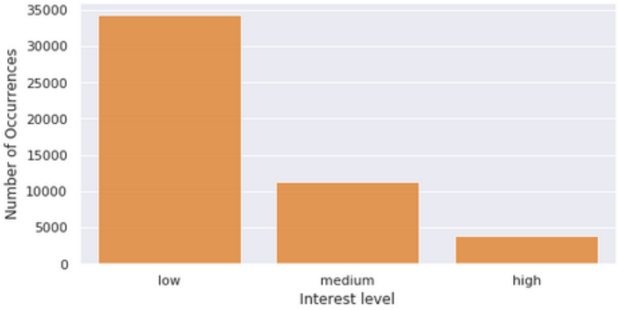


Fig. 4. Value Counts of target

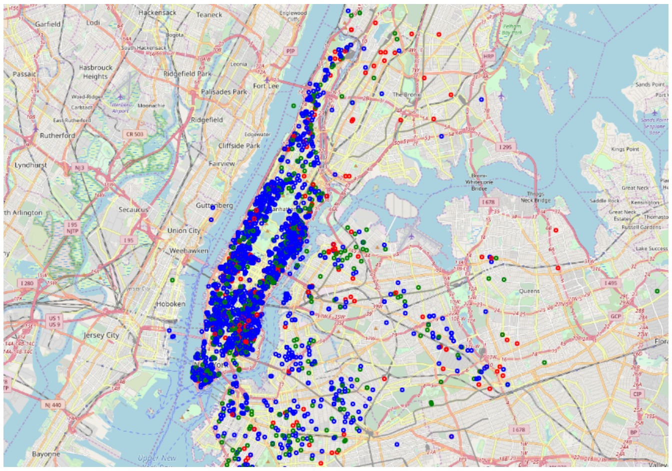


Fig. 5. Visual Representation of Interest lvl for NY

3.2 Machine Learning Algorithms

4 Results

5 Discussion

fd

moep moep

6 Conclusion