

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/>).

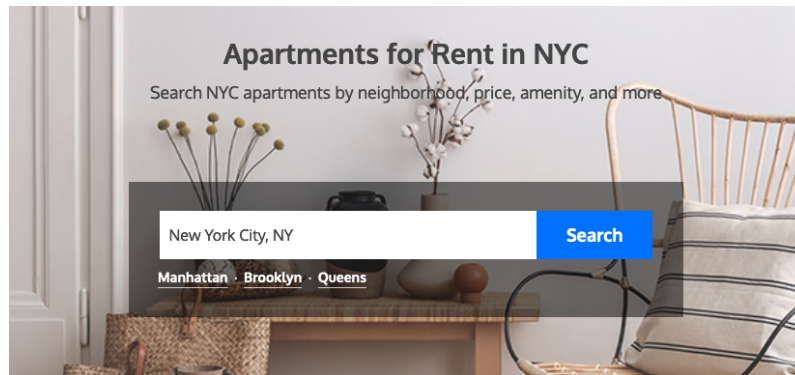


Fig. 1. Renthop Platform

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](https://www.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 decision 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.

1.4 Metrics

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

$$\text{logloss} = -\frac{1}{n} \sum_{i=1}^N \sum_{j=1}^M y_{ij} \log(p_{ij}) \quad (1)$$

2 Methodology

3 Results

4 Discussion

5 Conclusion