Compulsory Exercise 2: Predicting AirBnB-prices in Rome

Helle Villmones Haug Hjalmar Jacob Vinje Sanna Baug Warholm

20 April, 2023

Abstract

The purpose of the project is to use machine learning to predict the price of Airbnb nights in Rome XX TBA. The data set used is the rome_weekends.csv file from Kaggle (XX source). It was analyzed by X and X. Summarize key findings Why are findings important/novel?

Introduction: Scope and purpose of your project

Briefly introduce the broad idea of the problem or task that you chose and the respective data set that you use. This could be a classification task (e.g., predicting the species of flowers in the Iris data set) or a regression task (e.g., predicting the price of a house based on its features). Clearly define the scope of your project. What specific problem are you trying to solve? • Describe the source and give a referenc to where the data set is coming from. • Describe the purpose of your project is some more detail. What are the specific question that you want to answer in your project? Are you trying to find the best performing method or a good performing and light method that is easy to use? Who is your audience? Are you trying to discover the relations between different variables? Are you trying to find important predictors for your classification? Are you trying to draw some insightful understanding in a particular topic/domain? Importantly: Is the main purpose inference or prediction?

In this project, we aim to predict the total rental price of Airbnb listings in Rome during weekends, using a dataset called 'rome_weekends.csv' from Kaggle (https://www.kaggle.com/datasets/thedevastator/airbnb-prices-in-european-cities?select=rome_weekends.csv). This is a regression task, as we are predicting a continuous variable (total rental price) based on various features of the listings.

The dataset is obtained from a public platform, Kaggle, where users can upload, share, and collaborate on data science projects. The data contains information on Airbnb listings in Rome, such as room type, whether the room is shared or private, and if the host is a superhost. Our goal is to utilize this information to predict the total rental price of a given listing and evaluate the performance of two machine learning models.

The purpose of this project is two-fold:

1. To compare the performance of two different machine learning techniques: a deep learning model using the Keras library and a Random Forest model. We want to determine which model yields better prediction accuracy and whether there is a trade-off between performance and model complexity.
2. To uncover relationships between different variables and identify important predictors for the total rental price.

Our audience consists of data scientists, machine learning enthusiasts, and Airbnb stakeholders who are interested in gaining a deeper understanding of the factors affecting rental prices and improving price prediction models. The main purpose of this project is to predict the total rental price of Airbnb listings with high accuracy, while also uncovering relationships between variables that can contribute to a better understanding of the underlying factors that influence rental prices. In essence, our project focuses on both prediction and inference, with the goal of achieving high prediction accuracy and gaining insights into the determinants of rental prices.

Descriptive data analysis/statistics

Conduct descriptive data analysis to get an overview over your data (see this example for inspiration). Examples: • Report measures such as mean, median, range, standard deviation, and variance to describe the central tendency, variability, and distribution of a data set. • Scatter plots and correlation matrices across different variables and histograms of variables (see this example). • Box plots of variables.

Methods

Describe the methods that you are using in your project and explain in detail how you applied them. Depending on the task, these could include methods such as linear or logistic regression, decision trees, random forests, support vector machines etc. You should use several methods for your problem so that you can compare their performance. • Explain briefly how each method works, what its strengths and weaknesses are, both in general but also in the light of your project (how suitable is the method in your case?). 3 • Describe which hyperparameters are optimized for the methods (e.g., a shrinkage factor is a hyperparameter in Lasso). • Describe clearly how you evaluate the performance of the different models and methods (accuracy, MSE, misclassification error, CV error,. . .). Explain how each metric is calculated, and why it is a useful measure of model performance. • (optional) Consider and describe potential limitations of the methods and the chosen evaluation metrics.

Method 1 A deep learning model using ReLU

The ReLU activation function introduces non-linearity but it does so using linear segments.

The method is implemented with the following features:

- ReLU activation function
- The model weights are updated 50 times on the entire data set
- The batch size of the training data is 32
- 20
- MSE is the loss function used to measure the difference between the predicted and actual values
- MAE is used to measure the absolute difference between predicted and actual values
- The model has 4 hidden layers with 60, 100, 20 and 20 units

The Keras library functions have features such as an optimizer (“Adam optimizer”) and X, which in many cases contribute to improved results.

Method 2 Random Forest model

Random forest is a method that constructs several decision trees and then aggregates their predictions. It doesn’t make any assumptions about the underlying distribution of the data, but works by iteratively partitioning the feature space and fitting decision trees to each partition. Each decision tree makes a prediction, and the final prediction is the average or the mode of the predictions of all the trees.

Results and interpretation

We expected the random forest to be slightly worse than the ReLU based deep learning model. The results we got are XX.

Summary