Machine Learning Capstone Project

Price prediction in the AirBnB London Market

Capstone Proposal

Kjersti Mjøs Aase 21.11.18

Domain Background

From its start in a living room floor in San Fransisco 2007, Airbnb has gained an enormous popularity. AirBnb was valued at 38 billion usd in May 2018 *1)*

> ***Airbnb's Business Model At A Glance***
>
> *Airbnb operates an accommodation marketplace that allows people to list their available living spaces to be leased or rented by users looking for short-term lodging.*
>
> *Airbnb allows bookings at listings in more than 81,000 cities across191 countries*

When a host is to set a listing price they are referred to the hosting help page, and informed that the listing price is completely up to themselves. To support their decision they can search for listings in their city or neighborhood to get an ide of market prices.

https://www.airbnb.com/help/article/52/how-should-i-choose-my-listing-s-price?locale=en

Problem Statement

There is a huge potential in automating this process and optimizing the listing price. Especially if you are a host with multiple listings.

Beyond pricing is a provider of such a pricing software for vacation rentals. You are reccomended a price based on the health of the listing, with factors such as neighborhood, seasonality, local demand and more. But it is not clear which models and methods are operating behind the scenes.

That is why I want to develop a pricing recommender using historical data in the London airbnb market.

I see this regression problem as an excellent opportunity to demonstrate machine learning skills obtained during the Udacity Data Scientist for Enterprise nanodegree.

Dataset

Data on airbnb listings are available from Inside Airbnb and is sourced from publicly available information on the Airbnb site
http://insideairbnb.com/get-the-data.html

I have chosen to focus on the London market. There are 74153 listings with 96 features in the London dataset, along with 1.1 mill reviews and calendar of availability. There is a mixture of both numerical and categorical features. And it will be relevant to create dummy variables, f.ex. for the amenities.

Solution Statement

The intention is to predict a continuous variable - a recommended price. Thus the capstone is focused on developing a regression model.

Benchmark models

As a benchmark, I plan to use out-of-the-box Linear Regression model from sklearn.

Evaluation Metrics

Common evaluation metrics for regressor classifications are MAE (mean absolute error), MSE (mean squared error) and R2

Project Design

The project design follows a standard Machine Learning project structure:

- *Data exploration and data visualization*
    - Discuss features and statistics relevant to the problem.
    - Visualize important features to support the discussion
- *Data Preprocessing*
    - Discuss and handle missing data, skewness and outliers,
- *Feature Engineering*
    - Find relevant features and remove redundant features
    - Will also look into PCA and see if combining the most correlated features into components can reduce the size of the input.
    - Sklearn.feature selection may also be used for dimensionality reduction to boost performance on high-dimensional datasets. SelectKBest and SelectPercentile might be relevant *2)*
- *Model Testing and Selection*
    - Analyze multiple regression models based on the evaluation metrics.
        - Benchmark model Linear Regression
        - Ensemble models AdaBoost and Random Forest regressor
        - Xgboost
    - Choose the best model
- *Model Tuning*

    - Tune the selected algorithm to increase performance using gridsearch
- *Results and conclusions*
    - Discuss findings and possible improvements

Resources:

1. https://www.forbes.com/sites/greatspeculations/2018/05/11/as-a-rare-profitable-unicorn-airbnb-appears-to-be-worth-at-least-38-billion/#71347f262741
2. https://scikit-learn.org/stable/modules/feature_selection.html#univariate-feature-selection