Machine Learning Engineer Nanodegree

Capstone Proposal

Clement LEFEVRE July 10th 2017

Proposal

(approx. 2-3 pages)

Predict the occupancy rate of Airbnb appartments in Berlin.

Domain Background

(approx. 1-2 paragraphs)

In this section, provide brief details on the background information of the domain from which the project is proposed. Historical information relevant to the project should be included. It should be clear how or why a problem in the domain can or should be solved. Related academic research should be appropriately cited in this section, including why that research is relevant. Additionally, a discussion of your personal motivation for investigating a particular problem in the domain is encouraged but not required.

Sharing economy is booming thanks to the spread of internet. It can also redraw entire pieces of the economy, like taxi industry with Uber and Lyft. For this capstone project, i will work on AirBnB, and focus on the appartments rented in Berlin, where i live. The AirBnB system has experienced a significant growth in the last 4 years there, and in some ways contributed to the shortage of housing in Berlin. Due to the historically relatively low rents in Berlin, local inhabitants are very sensitive to this topic, which is perceived as a further syndrom of gentrification. As a consequence, the local press often raises the issue of Airbnb (New Berlin ban for Airbnb), AIRBNB vs. BERLIN.

For the persoanl aspect of this topic, one appartment in the building i live in is full-time rented via AirBnb, incurring disturbance such as loud music, smoking in the stairs, and broken glass.

Problem Statement

(approx. 1 paragraph)

In this section, clearly describe the problem that is to be solved. The problem described should be well defined and should have at least one relevant potential solution. Additionally, describe the problem thoroughly such that it is clear that the problem is quantifiable (the problem can be expressed in mathematical or logical terms) , measurable (the problem can be measured by some metric and clearly observed), and replicable (the problem can be reproduced and occurs more than once).

From the aspects highlighted above, how can we predict the occupancy rate of an AirBnB appartment given the data publicly available ? We can define the occupancy rate as the proportion of booked nights over the next 6 months.

Datasets and Inputs

(approx. 2-3 paragraphs)

In this section, the dataset(s) and/or input(s) being considered for the project should be thoroughly described, such as how they relate to the problem and why they should be used. Information such as how the dataset or input is (was) obtained, and the characteristics of the dataset or input, should be included with relevant references and citations as necessary It should be clear how the dataset(s) or input(s) will be used in the project and whether their use is appropriate given the context of the problem.

As the AirBnb data are not directly available, i used the dataset from insideAirBnB which scraps for some cities the entire offers and make those datasets available.

The dataset for each city consists in a list of appartments with all features available using the regular AirBnB booking website, including the reviews of the past guests.

Solution Statement

(approx. 1 paragraph)

In this section, clearly describe a solution to the problem. The solution should be applicable to the project domain and appropriate for the dataset(s) or input(s) given. Additionally, describe the solution thoroughly such that it is clear that the solution is quantifiable (the solution can be expressed in mathematical or logical terms) , measurable (the solution can be measured by some metric and clearly observed), and replicable (the solution can be reproduced and occurs more than once).

Benchmark Model

(approximately 1-2 paragraphs)

In this section, provide the details for a benchmark model or result that relates to the domain, problem statement, and intended solution. Ideally, the benchmark model or result contextualizes existing methods or known information in the domain and problem given, which could then be objectively compared to the solution. Describe how the benchmark model or result is measurable (can be measured by some metric and clearly observed) with thorough detail.

Such predictive tool are already available as commercial tool AirBnb Investment Explorer, but they lack the necessary transparency to understand the logic behind it.

Evaluation Metrics

(approx. 1-2 paragraphs)

In this section, propose at least one evaluation metric that can be used to quantify the performance of both the benchmark model and the solution model. The evaluation metric(s) you propose should be appropriate given the context of the data, the problem statement, and the intended solution. Describe how the evaluation metric(s) are derived and provide an example of their mathematical representations (if applicable). Complex evaluation metrics should be clearly defined and quantifiable (can be expressed in mathematical or logical terms).

As we are in the context of the regression problem, i will use the Root Mean Squared Error metrics (RMSE) as the main metrics.

The RMSE is defined as : The translation $[\![e]\!]$ given by

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}$$

where y_i is the real occupancy rate and $\bar{y}_i$the predicted one.

Project Design

(approx. 1 page)

In this final section, summarize a theoretical workflow for approaching a solution given the problem. Provide thorough discussion for what strategies you may consider employing, what analysis of the data might be required before being used, or which algorithms will be considered for your implementation. The workflow and discussion that you provide should align with the qualities of the previous sections. Additionally, you are encouraged to include small visualizations, pseudocode, or diagrams to aid in describing the project design, but it is not required. The discussion should clearly outline your intended workflow of the capstone project.

The first step will consist in transforming the text reviews into usable features. I will use TFIDF and filter on the most relevant words.

Then i will implement differents regression algorithm (linear, XGBoost) and also give a try with Neural Networks using the keras wrapper for tensorflow.

Finally, i will use the stacking technique to get the best predictor as a combination of differents models.

Before submitting your proposal, ask yourself. . .

- Does the proposal you have written follow a well-organized structure similar to that of the project template?
- Is each section (particularly **Solution Statement** and **Project Design**) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?
- Would the intended audience of your project be able to understand your proposal?
- Have you properly proofread your proposal to assure there are minimal grammatical and spelling mistakes?
- Are all the resources used for this project correctly cited and referenced?