Predicting Virginia Beach Resident Satisfaction Rate with Machine Learning

A Preprint

Alan Zhai
Computer Science and Environmental Science
University of Virginia
asz9qm@virginia.edu

Siyeon (Shaun) Kim
Computer Science and Statistics
University of Virginia
sk5ps@virginia.edu

Jaspreet Ranjit
Department of Computer Science
University of Virginia
jr4fs@virginia.edu

July 30, 2020

1 Abstract

The goal of this project is to leverage machine learning techniques to produce a model that predicts overall life satisfaction for residents of Virginia Beach using various aspects of the city such as public transportation, commodities, and other properties relevant to city planning. The primary focus of this research is Virginia Beach, but the goal is to produce a framework for a model that can be generalized and used in other cities to predict life satisfaction rates for residents. The data from the survey was standardized and pre-processed to convert categorical variables to numerical values. The problem of predicting life satisfaction was modeled with regression and random forest models. Demographic features of residents were also introduced to serve as another avenue of exploration in predicting life satisfaction rates.

2 Motivation

The primary focus of the project is the overall satisfaction score of the residents of Virginia Beach predicted on the basis of the characteristics of the city. As students who are interested in urban planning, it is important to know what aspects and design features of a city are working well and what areas could use improvement with the overall goal being to increase citizen satisfaction. Furthermore, current city planners use common statistical techniques to model the data and draw conclusions. Although this can be useful to visualize the data, it does not provide much insight or information as to what factors contribute the most to a resident's living experience. As a result, it would be beneficial to explore and develop a model that could give city planners more insight into which aspects of the city need to be improved, or which aspects should be more heavily considered in planning new cities.
Another use case of the proposed model is for prospective home buyers who are looking to move to another location. The model can provide guidance as to which city will provide better quality of life and help individuals make more informed decisions. In macro-scale, although the scope of the data is limited to Virginia Beach, the proposed model can be extended to other cities who collect similar citizen satisfaction data. For this specific motivation, the demographic data will be used as a predictor from the survey result to tailor one's satisfaction score more personally.

3 Method

Prior to discussing what models were used for this preliminary experiment, it is crucial to go over how the label - the overall satisfaction score - is calculated in the study. The overall satisfaction score is a summation of encoded value of the survey result from Q2-Q36. These sets of survey questions ask a subject how satisfied he or she is with specific aspect of Virginia Beach. For "very dissatisfied," the score is 1; for "dissatisfied," the score is 2; for "satisfied," the score is 3; for "very satisfied," the score is 4. The higher overall satisfaction score indicates that the observation is more satisfied with the living conditions of the Virginia Beach. In order to predict the overall satisfaction score, we used the following methods for our preliminary experiment:

1. ***Linear Regression without Demographic Data***
   The main reason the linear regression model was chosen to predict the satisfaction score is because this is the most standard model for prediction and always a good first step towards building a more complex model. The main difficulty expected with this model is that since the independent variables are mostly categorical; we do not expect linear regression to be the most accurate model for that reason.

2. ***Linear Regression with Demographic Data***
   The second part of the goal for this experiment was to test with demographic data for each of the residents. TO provide a more personalized prediction of the satisfaction score, the demographic data was incorporated as features. Demographic data included gender, income, education level, number of years that the residents lived in Virginia Beach, etc. This data, again, was modeled with the linear regresssion model.

4 Preliminary Experiments

Model

The preliminary experiment included data preprocessing and cleaning, and testing the linear regression model. Since the data consisted of categorical variables and the task described is defined as a regression problem, the predictors were one hot encoded to convert the categories to numbers. After scaling the processed data, the linear regression model from the scikit learn built in library was used to predict the satisfaction score.

Initial Results

As described in the 'Method', the linear regression model was used to predict satisfaction score. There were two variations of this preliminary experiment: a prediction including demographic data about the residents and a prediction without the demographic data. The root mean square error (rmse) for the experiment that included demographic data in the predictors was 0.13158 and rmse for the experiment that did not include the demographic data was 0.13999. Given the context of the problem, this error is too high. Given that the mean and standard deviation are 5.595749 and 0.169593 respectively, the error spans almost one standard deviation and thus does not serve as a good predictor of the satisfaction score. The linear regression model is not complex enough for the problem presented and the data does not fit well to the model: both the training and testing error are 0.125 and 0.1315, so additional models will need to be investigated.

The link to the Google colab can be found here: `https://colab.research.google.com/drive/1L4G7MNOO2mWcffCC-MchIl0Uy7b3AADJ`

5 Next Steps

Since the existing model did not necessarily perform well given the context of the problem, we plan to do more extensive preprocessing and build more complex models. First, we plan to preprocess features differently. After looking at the data granularly, we realized that some of the independent data could be scaled instead of 0,1 coded. That will is the initial thing we plan to do. Afterwards, we plan to build more complex models like Random Forest, SGDRegressor and Polynomial Regression. For the regressions, we will also attempt to scale the label, putting sqrt or log around the regression equation until we get the lower rmse and a better performance. Since, we have a lot of independent variables, we will use Lasso regularization method to feature eliminate some of the redundant features. We hope that this regularization will drastically improve the prediction.

6 Contributions

Every member of the team contributed to finding the appropriate data set and developing our main interest for the project. Alan pre-processed and cleaned the data and initiated building models for this check point. The report is mainly drafted and finalized by Shaun and Jaspreet. The rest of the members will implement at least one statistical method or variation of the current model for the final report.