Inferential relationships between weather characteristics and cost of dwelling in the United States

Alex Johnson
Undergraduate Student, Department of Computer Science and Electrical Engineering
Brigham Young University – Idaho
Rexburg, ID, USA
Email: joh13118@byui.edu

1. Introduction

Real-world predictive models are often challengingly complex, and sometimes mathematically overwhelming, to implement in search of reliable decision or insight mechanisms. The advent of modern machine learning techniques has introduced data scientists and engineers to newer ways of finding fascinating correlative indicators that have successfully hidden themselves away in datasets that have sat untouched for decades.

This semester-long research project is an attempt to capture some of the excitement about machine learning through mining complex datasets to uncover interesting inferential data.

1.1 Objective and scope

This project seeks to make inferences about the cost of dwelling in different areas of the United States based on weather data. Many publicly available datasets contain enough historical observations of both weather patterns and dwelling costs to train a machine learning algorithm with a healthy ratio of training to testing data.

1.2 Datasets used

Because it is difficult to find a single dataset that contains both historical weather and cost of dwelling data that also has a sufficient number of observations, two datasets were obtained that provide many observations concerning each topic. These datasets are the *Zillow Rent Index* and the *Historical Hourly Weather Data*, each of which were obtained publically from www.kaggle.com. The data populating the *Zillow Rent Index* was obtained in JSON format from Zillow's public API, and the data in *Historical Hourly Weather Data* was obtained in a similar manner from www.openweathermap.org.

	City Code	City	Metro	County	State	Population Rank	November 2010	December 2010	January 2011
0	6181	New York	New York	Queens	NY	1	NaN	NaN	NaN
1	12447	Los Angeles	Los Angeles	Los Angeles	CA	2	2184.0	2184.0	2183.0
2	17426	Chicago	Chicago	Cook	IL	3	1563.0	1555.0	1547.0
3	39051	Houston	Houston	Harris	TX	4	1198.0	1199.0	1199.0
4	13271	Philadelphia	Philadelphia	Philadelphia	PA	5	1092.0	1099.0	1094.0
...	...	...	...	...	...	...	...	...	...
13126	397405	Highland Township	Gettysburg	Adams	PA	13127	1280.0	1280.0	1284.0
13127	398292	Town of Wrightstown	Green Bay	Brown	WI	13128	639.0	650.0	668.0
13128	398343	Urbana	Corning	Steuben	NY	13129	1433.0	1431.0	1437.0
13129	398839	Angels	NaN	Calaveras	CA	13130	1516.0	1529.0	1529.0
13130	737788	Lebanon Borough	New York	Hunterdon	NJ	13131	1759.0	1784.0	1787.0

13131 rows × 81 columns

Figure 1. Example data from *Zillow Rent Index.*

1.3 Outcomes

Several interesting outcomes were observed from experimentation with the data:

- Several types of algorithms performed well with the data, even though the objective was essentially regression-based.
- While weather proved to be a reasonably accurate indicator of the general range in the cost of dwelling, there also appear to be latent factors that are not represented in the data.
- Support Vector Machine-based models made interesting inferences on dimensions not explicitly available in the data. It is possible that some latent factors were

uncovered by Eigen dimensionality analysis performed during experimentation with SVMs.

- Of the weather features present, humidity seemed to play a "fine-tuning" factor in the range of cost of dwelling. E.g., although higher average temperatures alone could indicate a higher cost of dwelling, the additional presence of high average humidity often negatively affected the expected cost.

2. Data preparation

The *Zillow Rent* Index and the *Historical Hourly Weather* datasets contain subjective data that form an intersecting set. Observations containing these shared subjects were used to form new feature-target observations to determine if a machine-learning algorithm could be trained to predict costs of dwelling given weather data.

2.1 Missing values

Although each dataset contained a significant number of missing values, the features taken from each set and used for machine learning were mostly present throughout each observation. Missing values were either imputed on the mean where the feature did not deviate significantly from the mean, or on the mode where the feature was categorical in nature.

2.2 Data complexity and standardization

The datasets contained multiple types of data, including integer categorical and text categorical, as well as discrete numeric and continuous numeric. Any ordinal features were not used for training the machine learning algorithms.

Because the observations in both datasets consisted of highly-fine, time series-based data, several resamplings were performed where the data was grouped by hour, month, and year, for example.

Numeric data standardization occurred in several forms, namely through scaling, label encoding, and one-hot encoding, depending on the type of data in question.

Once the essential features were chosen by allowing XGBoost to perform feature analysis and F-Scoring, the datasets were distilled together into the most important features. This helped the model perform significantly better with regards to its performance metrics, such as Root-Means-Squared-Error. The combined dataset that performed the best consisted of five features and one target.

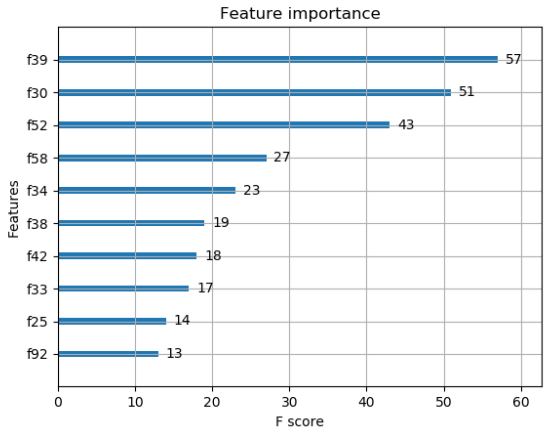

Figure 2. XGBoost feature analysis.

3. Algorithms

Several algorithms were experimented with while trying to find the best model for the available data.

3.1 SKLearn Support Vector Machine

Both datasets contained a large number of features, and this dimensionality likely gave way to latent factors within the data. The SVMs used helped the data show that the population of a given geographical location was probably a large factor in determining the cost of dwelling. However, consistent performance with regression analysis was lacking.

3.2 SKLearn Decision Tree

Decision trees are usually efficient at performing either classification or regression tasks if provided with sound data. However, the single decision tree

algorithms employed also struggled with regression analysis in these particular datasets.

3.3 XGBoost Regressor

As with most other tasks, XGBoost proved the most efficient algorithm out of those tested against the data. XGBoost was able to outperform any other algorithm while minimizing its RMSE metrics lower than its competitors.

For these reasons, XGBoost was chosen to perform the principal analysis on the finalized datasets.

```
In [21]: rmse = np.sqrt(mean_squared_error(y_test, predictions))
         print(rmse)

         262.30571
```

Figure 3. Sample RMSE score during XGBoost training.

4. Results

The XGBoost Regressor algorithm uncovered exciting relationships in the data. Positive correlations were discovered between costs of dwelling and"desirable" weather characteristics, such as temperature, while humidity and pressure were often tied to lower costs of housing. XGBoost Regressor's proficiency at determining the general range of dwelling cost given the essential features is reflected in its respectably-low RMSE curve throughout various experiments.

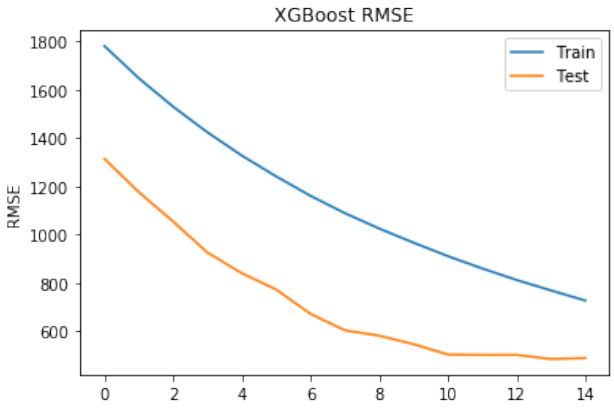

Figure 4. XGBoost Regressor's general validation curve.

When fitting the model on only the most important features, the trees generated by XGBoost Regressor tended towards five to six levels. When limiting the model to 4 levels or less, performance dropped significantly.

Figure 5. XGBoost Regressor tree with 5 levels (3 shown).

5. Conclusions

5.1 Business and stakeholder considerations

Accurate regression models can help facilitate important decisions at every level of society. Businesses and stakeholders with interest in real estate and property brokerage must be well-informed of the many factors that can affect their respective markets. Businesses seeking to expand their property prospection across regions, or even within the surrounding area, may find that the impact weather and climate have on housing costs is worth consideration. Beyond the direct correlations found in the data, latent factors such as relative population size and absolute population index could also help invested parties make better business decisions.

5.2 Ethical considerations and limitations

As with any machine-learning endeavor, special attention must be given to any potential ethical considerations. Publically-available data concerning housing may contain residents' private information or other data whose appearance on the Internet may be questionable. Fortunately, all of the data contained in the datasets used in this study was obtained from reputable sources using legitimate collection techniques.

The results obtained from this project are to be treated lightly due to the relatively limited number

of experiments performed on vast amounts of data. While inferential correlations may have been discovered, by no means does the data suggest that any causality has been found. The observations made are the results of educational experimentation with cutting-edge analysis technologies.

6. Lessons learned

Statistical analysis of complex topics is difficult and time-consuming; the results obtained from this project are likely rudimentary compared to those that could be produced by an expert. However, this project presented a vast number of opportunities to learn best practices and further develop foundational analysis techniques.

If I started this project over again, I would place more importance on Principal Component Analysis and ensure that as many latent factors could be uncovered as possible; it is almost certain that a latent factor hid behind every correlation discovered during the project. I would also make better use of data preparation techniques to provide more finely-tuned data to the algorithms chosen.

This project has been a great learning experience and has provided much motivation to continue developing my skills and knowledge in machine learning.