Readme.md

KC Homebuilding 301

Authors: Peter Vuong |Brian Reynolds |Will Norton

Business Understanding & Problem

Our stakeholder, Howard S Wright, is a well-known construction company that is prevalent within the King County area. Howard S Wright wants to know what features should they focus on in order to maximize their profit when building a new home in King County. Some factors that we intitially thought they could focus on were location, square footage, number of bedrooms/bathrooms, and quality of the home.

Data & Methods

We utilized King County housing data from 2014 to 2015, as well as demographics data we acquired online from King County's census data. We utilized a GIS software in order to retrieve the specific data that we needed. We dropped dupplicate entries and data columns not relevant to our stakeholder, such as year renovated (yr_renovated). In addition we engineered some features, Bed Bath Ratio, Footprint, and Sqaure Feet per Bedroom to use in modeling. Finally, we used the GIS software & demographics information to determine what zipcodes were in each court district of King County and binned the zipcodes into their respective regions.

Modeling

Using the dataset, we then embarked in an iterative modeling process using multiple linear regression.

First Model & Subsequent Iterations

For our first model, we kept it simple, removing categorical variables and limiting the number of features in order to alleviate problems with multicollinearity. Unfortunately, this model using square foot living, floors, and bedrooms proved to be a mediocre performer, only accounting for 50.5% of the variation in price (r^2 .505).

In the following models, we used different groups of the continuos variables, including those the we engineered, in an attempt to improve our r-squared and reduce error. Unfortunately, this proved fruitless resulting in models with worse r-squared and still very high error metrics.

Categorical Variables, Final Model, and Error Metrics

Next we turned to the categorical variables of grade and region, per our binnned zipcodes. Our fourth model took the kitchen sink approach with eight variables (bedrooms, bathrooms, square foot living, square foot lot, floors, year built, zipcode, grade) and gave our best r-squared (.702). However we were concerned that this would be more complex than our stakeholder would be interested in.

Our final model narrowed things down to just three variables: square foot living, zipcode, and grade. This resulted in a model that performed signicantly better than our first, .680 vs .505, along with being simpler than our fourth model.

The good news is that our final model performed very similary on both our training and test data, indicating that it was not overfit. Unfortunately, the root mean squared error remained stubbornly throughout the modeling process, with our final model having a RMSE of 206022. Obviously, an average error of over $200,000 for the predicted price of a home is not ideal. Further research into more advanced modeling techniques is needed to produce a less error prone model.

Conclusions

Our business stakeholder, Howard S Wright, should focus on three important factors while building homes:

1. Square footage
2. Location, with the West region fetching the highest sale prices
3. Quality of construction

The second and third recommendations come with the caveat that there are a limited number of very large and very high quality homes sold in our dataset. Any home building plans should consider the volume of homes sold.

Future Research

There are three main areas where we would like to investigate further:

1. Look at home demand by region & buyer demographics, including factors like population growth & income.
2. Examine the impact of using eco-friendly building techniques on sale price and demand, as that appears to be a popular feature in the Pacific Northwest

3. Analysis of construction costs to sale price in order to maximize profits. As an example, our modeling process already found that thenumber of floors did not have a significant impact on price and two story homes are generally cheaper to build.

Repository Structure

```
.
├── data
├── deliverables
├── images
├── .gitignore
├── Project 2 Final Notebook.ipynb
└── README.md
```

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 3


NaavSD


petercvuong

Noptov

Languages

- **Jupyter Notebook** 100.0%