tjkyner removed .gslides file ... 3 minutes ago 30

data	initial commit	12 days ago
images	finished presentation	7 minutes ago
submissions	removed .gslides file	3 minutes ago
.gitignore	fixed .gslides in gitignore	4 minutes ago
LICENSE	Initial commit	13 days ago
README.md	changed wording in business problem	2 hours ago
notebook.ipynb	finished presentation	7 minutes ago

README.md

King County Housing Data Project

Student name: T.J. Kyner
Student pace: Full time
Instructor: Abhineet Kulkarni
Cohort: 040521

Project Overview

Business Problem

The ability to accurately appraise a house is of critical importance for a variety of stakeholders. In addition to buyers and sellers, which each have their own interests in finding the fair market price of a house, other entities such as municipalities benefit from such insight as well. Given that property taxes provide a large portion of tax revenues for municipalities, having an accurate prediction model for house prices can play a key role in efficient financial planning and budgeting. The goal of this project is to provide such a prediction model for the benefit of municipalities in King County, Washington.

Repository Structure

```
├── data/
│   ├── kc_house_data.csv  # Data on house sales in King County
│   └── column_names.md    # Descriptions of the columns in kc_house_data.csv
├── images/                # Contains exported images of plots used in the presentation
├── submissions/           # Contains files used for the project submissions
├── .gitignore
├── LICENSE
└── README.md
```

Exploratory Data Analysis

While most of the 21 columns of data available are fairly self-explanatory, some require a bit more explanation. Along with the original dataset, some metadata on the column names was also provided. Brief descriptions of each column are as follows:

Column	Description

About

This repository contains my phase 2 project for Flatiron School's data science program.

Readme

GPL-3.0 License

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

Jupyter Notebook 100.0%

Column	Description
`id`	unique identified for a house
`date`	house was sold
`price`	is prediction target
`bedrooms`	of Bedrooms/House
`bathrooms`	of bathrooms/bedrooms
`sqft_living`	footage of the home
`sqft_lot`	footage of the lot
`floors`	floors (levels) in house
`waterfront`	House which has a view to a waterfront
`view`	Has been viewed
`condition`	How good the condition is (Overall)
`grade`	overall grade given to the housing unit, based on King County grading system
`sqft_above`	square footage of house apart from basement
`sqft_basement`	square footage of the basement
`yr_built`	Built Year
`yr_renovated`	Year when house was renovated
`zipcode`	zip
`lat`	Latitude coordinate
`long`	Longitude coordinate
`sqft_living15`	The square footage of interior housing living space for the nearest 15 neighbors
`sqft_lot15`	The square footage of the land lots of the nearest 15 neighbors

Expanded definitions for certain columns, such as `condition` and `grade`, can be found within King County's Residential Glossary of Terms.

While most of the provided descriptions seem logical, I do have a concern regarding the `view` column. It seems more logical for this column to be referring to some sort of grading scale for the view available from the house rather than if it has been "viewed" (by whom? for what purpose?). This idea is further supported by "Views" being defined in King County's Condo Glossary of Terms as follows:

> For each classification will display blank for no view or "Fair", "Average", "Good" or "Excellent" to reflect the quality of view for that unit

This section of the notebook involved looking at summary statistics and generating basic visualizations in order to make observations about any issues or concerns that warranted further investigation or correction within the preprocessing section. Some of the information visualized includes:

- Scatter plots for each variable versus `price`
- Correlation heat map
- A rough map by using the latitude and longitude data in a scatter plot

Data Preprocessing

Based on the observations made in the EDA section, the following list represents the goals for preprocessing the data before moving on to creating the baseline and subsequent prediction models:

1. Drop the `id` column
2. Investigate splitting the `date` column into two columns containing the month and year
3. Convert the `sqft_basement` column to an integer and handle placeholder values
4. Drop the `yr_renovated` column

5. Handle missing values in `waterfront` and `view`
6. Handle multicollinearity between highly correlated columns

Model Building

The model building process is iterative. Beginning with a baseline model, each subsequent model built upon the previous and applied a new adjustment / transformation (with the exception of model 6 which built off of model 4).

1. Baseline Model

- No adjustments beyond those made in the preprocessing section that apply to all models

2. Removing Outliers

- Removed any houses with more than eight bedrooms
- Dropped a 2,300 sqft house with only a half bath

3. Categorical Variables

The following variables represent categorical data and were turned into dummy variables:

- `waterfront`
- `view`
- `condition`
- `grade`
- `zipcode`
- `month`

4. Log Transformation

The following variables represent continuous data that exhibited skewness and were able to have a log transformation applied (all values were positive and non-zero):

- `price`
- `bedrooms`
- `bathrooms`
- `sqft_lot`
- `sqft_above`
- `yr_built`
- `lat`
- `sqft_living15`
- `sqft_lot15`

Achieved the highest R^2 value at approximately 88.5%.

5. Scaling

The `RobustScaler` from `sklearn.preprocessing` module was applied which removes the median and scales the data according to the IQR. This particular scaled was chosen due to its decreased sensitivity to outliers versus other popular scalers such as the `MinMaxScaler`. This adjustment did not change the R^2 value but significantly reduced the interpretability of the model.

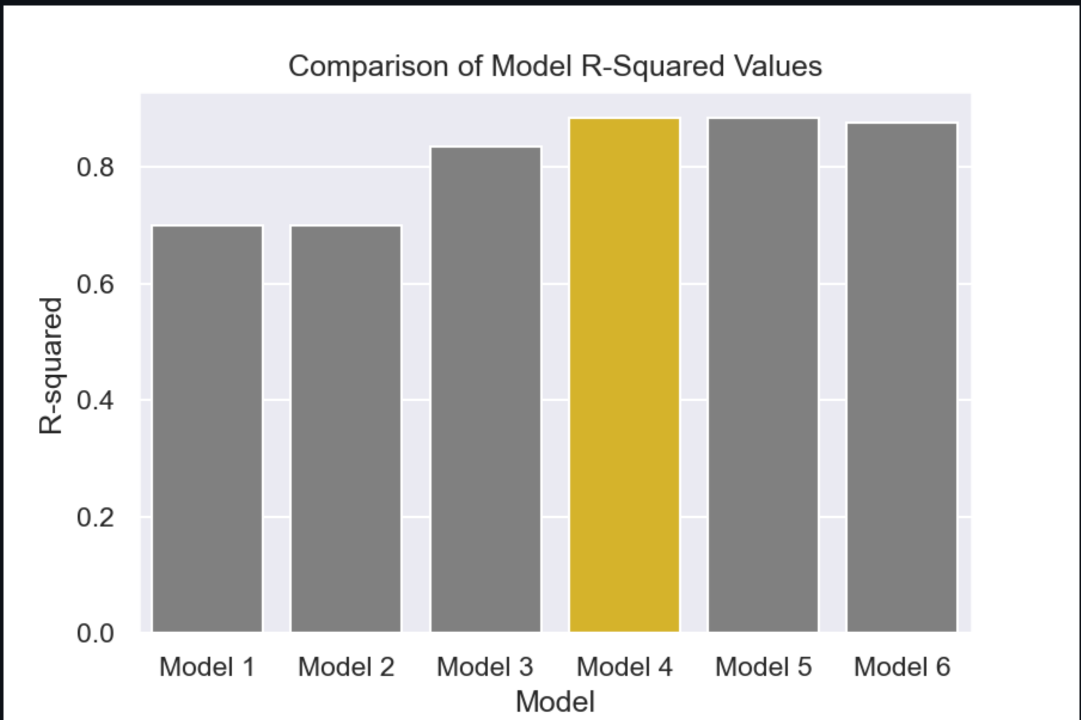
6. Dropping Non-Significant Variables

This iteration picked up from model 4 instead of model 5. The previous adjustments / transformations led to a number of different independent variables having non-significant p-values. This iteration dropped those variables in an attempt to make a leaner model.

Model Comparisons

Model 4 was selected as the final model as a result of having the highest R^2 value achieved and being more interpretable than Model 5.

Model	R-squared
Model 1	0.699828
Model 2	0.700233
Model 3	0.835001
Model 4	0.885039
Model 5	0.885039
Model 6	0.874982



Conclusion

Results

The fourth model, which removes outliers, includes dummy variables for categorical data, and log transforms continuous data, was the best performing model. This model explains approximately 88.5% of the variations in price for houses in the dataset. Some of the most impactful variables include:

- Being located in zip code 98039 (Medina, WA)
- Having a waterfront property
- Having higher rated `condition` and `grade`
- Being further north (higher latitude)

While not perfect, this model has the potential to be a useful tool for municipalities seeking a better estimate of future tax revenues. Instead of relying on the results of infrequent and costly appraisals for an estimate of taxable value, this model can provide a decently accurate estimate in a short amount of time.

Next Steps

There are many additional ways in which this model can be improved upon over time.

- Further iteration on the model to test for non-additive interactions and various other transformations
- A direct incorporation of an adjustment to the predicted house values to derive the estimated taxable value
- Enhanced location data that includes items such as proximity to amenities and walkability
- Inclusion of macroeconomic variables such as mortgage rates, new constructions, bank lending conditions, etc.