main 1 branch 0 tags Go to file Add file Code

schahmatist added pdf and html - FINAL 83ba35c 10 minutes ago 34 commits

analysis_and_regression	Second Cleanup	3 hours ago
code	Final Cleanup	3 hours ago
data	Initial Setup	8 days ago
images	Second Cleanup	3 hours ago
pdf	added pdf and html	11 minutes ago
tex2pdf.-0929dff1d79813bd	added pdf and html	11 minutes ago
.gitignore	Final Cleanup	3 hours ago
README.md	Overview update for Readme.md	2 hours ago
for_pdf_non-technical.ipynb	added pdf and html	11 minutes ago
non-technical.ipynb	added pdf and html	11 minutes ago
presentation.html	added pdf and html	11 minutes ago
technical.ipynb	Second Cleanup	3 hours ago

README.md

Phase 2 Project: Technical Presentation of Price Predictor

Overview

- The goal of this project is to develop a predictive model for house pricing in King County.
- The model will estimate how the features of a house will affect its price.
- The price estimation tool may be benefitial for Real Estates Agencies and Developers, as well as individual sellers and buyers.

Challenges

- Determining how multiple features work individually and together
- Quantifying joined features effect
- Building a predictive model
- Building a front end for a customer

Solution

- Analyzing 2014-2015 dataset with past sales
- Identifying individual and joined factors.
- Prepairing features for the model
- Calculate all the features coefficients
- Testing the results

Data

King County house sales dataset contains:

- details for 22,000 sold houses
- final sales prices

All the data is from 2014-2015

Features Identified

Main Features:

- House Sq footage
- Grade of design and materials quality
- Zipcode
- Waterfront
- View

Additional Features:

- Lot size
- Basement
- House Age

Only marginal effect from:

- Renovation, number of bedrooms, bathrooms, and floors

more on feature analysis - see "analysis_and_regression/Investigation of Features.ipynb"

Initial Data Load and Cleaning

- **Loaded the "kc_house_data.csv" using "initial_data_prep.py"**
- **filled or removed rows with missing properties**
- **Construction Grade 3-5 (below the acceptable code) were removed**
- **Out of 22,000 rows 20,880 were used in the model**

Data Modeling

An iterative approach to data modeling

- Calculating Efficiency for basic features
- Prepairing model features
- Training multiple models
- Chosing the most efficient model
- Testing against different subset of data

Steps:

- Prepared data for modeling using custom "transform_data" function (see functions_v1.4.py)
- Created/trained model using statsmodels.OLS
- Made sure r-square is higher than 80%

In addition to automatic -sklearn- methods, custom functions were created to manually get all the coefficients from statsmodels OLS and calculate the linear slopes formula

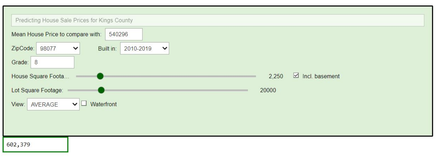
- used custom function "calcuate_price" and "get_coeff" to get coefficients from ols model (see functions_v1.4.py)

Creating UI forms

Building a Front End Tool:

- ipywidgets were used to create custom ui forms (Build_Forms_v1.4.py)

* custom calculate_price function was linked to the input/output of the ui



Testing

We made sure the tool works as expected:

- Multiple comparissons of predicted data against the actual data
- Predicted price is within 90-110% of actual price (houses newer than 1980)
- Predicted price is within 87-113% of actual price (houses older than 1980)

More deails about regression testing in "analysis_and_regression/Regression Tests.ipynb"

Conclusions

Considerations and Limitations:

- The tool can be effective to estimate base price for known features
- In the future a model should be re-trained with more up-to-date data
- The presented prototype will be greatly improved by more advanced modeling