Dmvinedata / dsc-phase-2-project-v2-3 Public

forked from learn-co-curriculum/dsc-phase-2-project-v2-3

Project 2_House_Prediction

☆ 1 star 263 forks

☆ Star ▾ Notifications

Code Pull requests Actions Projects Security Insights

final_package ▾ Go to file

This branch is 20 commits ahead of learn-co-curriculum:main.

Dmvinedata add student ... 36 minutes ago 36

View code

README.md

King County Home Value Predictions

Presentation Link Jupyter Notebook Link

Author: Deztany Jackson

Intial Date: December 2022

Overview

Real Estate agents in King County, Seattle are evaluating the neighborhoods to encourage current home owners of he benefits of improving and upgrading their property value. Housing data from King County was used to develop linear regressions models to support future price prediction.

--

Business Understanding

The primary stakeholders are real estate agents because of their wide use cases, network, domain knowledge and their incentive for home owners to increase their property value. They can also use this for getting a jump start on marketing to potential home buyers. The same predictions could be useful for the homeowners, potential buyers and even those in the remodeling and construction business. Because of their connection and real estate agents are able to influence a larger community's property value which as greater impact than convinving individual homeowners. The area attracts new implants from tech jobs. A great number of these people (as singles or families) may be looking to buy or rent.

This model is used as an intial model supporting course predictions. The main attributes used to support model prediciton are: Condition and Grade. The main attributes used to support model creation are:

'bedrooms', 'bathrooms', 'sqft_living', 'sqft_lot', 'floors','waterfront', 'yr_built', 'zipcode', 'cond_num', 'grade_num'

I would not guarantee the price predicions are 100% accurate, but will be useful to support general predictions. The model explains 67% of the variation. This means there are things that the model still doesn account for in making predictions.

Phase 1 Project Description, 2022

Data Understanding

This project uses the King County House Sales dataset (from GitHub project repo). This can be found in several locations: Git Hub Data & Kaggle.

The data is in the format of "csv". The initial data used in the modeling will start with 20 of possible attributes. As modeling progressing certain attributes (features) will be processed, transformed and possibly removed.

The imported data is the entire initial dataset. This will be scoped down before the intial understanding to support the the core modeling needs with the limited time available. Starting with the entire dataset could potentially support a more accurate model with the increase of the attributes to choose from. We will settle for good enough with the dataset we have.

Initial Features Used

- **King County Table**
  - Rows: 21597
  - Features: 11
  - "id" is not a feature used in the modeling

Cardinal Numbers

- ***price - (Target Variable)***
  - Description: Price is the amount of the house in context of the current attributes.
  - Type: Float Number
  - Expectation/Comment: Price will be out target variable. We will want to see after developing a solid model the varying of attributes would effect price. The price difference after an "upgrading" the home is also needed.
- **bedrooms**
  - Description: Number of bedrooms for the given home
  - Type: Int Number
  - Expectation/Comment:Will evaluate how well this impacts the model.
- **bathrooms**
  - Description: Number of bathrooms for the given home
  - Type: Int Number
  - Expectation/Comment: It has the second highest correlation value against price. These are.25, .5 and .75 are use in addition to whole numbers. .25-Sink; .5-Sink and Toilet; .75-Sink, Toilet and Shower,/Bath; 1 - Everything
- **sqft_living**
  - Description: The size of the livable space in the house
  - Type: Int Number
  - Expectation/Comment:We will assume larger will amount gather more money.
- **sqft_lot**
  - Description: The size of the lot

- Type: Int Number
- Expectation/Comment: We will assume larger will gather more money.

- **floors**
  - Description:'he Number of Floors.
  - Type:float Number
  - Expectation/Comment: We are not using the other attributes that compliment the number of floors.Sometimes one floor could be desirable. It would be hard to understand the house is architected with floors. It could be one floor and a basement, or two floors and no basement.

Nominal Numbers

- **yr_built**
  - Description: Year when house was built
  - Type: Int Number
  - Expectation/Comment: It could give context to the grade and condition. This may be useful for changing variables for prediction purposes
- **zipcode**
  - Description: ZIP Code used by the United States Postal Service
  - Type: Int Number
  - Expectation/Comment: The Zipcode will be there to help add to the business case. It can be helpful in creating collection of houses to focus on.

Categorical Objects

- **waterfront**
  - Description: Whether the house is on a waterfront
  - Type: Object String
  - Expectation/Comment: There isn't any information on what type of water. For those that are Null, we will fill Null with Unknown
- **condition**
  - Description: How good the overall condition of the house is. Related to maintenance of house.
  - Type: Intial Object String (Transformed to Int) - 5 Values
  - Expectation/Comment: This paired with the grade may the the top useful in changing for prediction. Will be transformed into number from string later in model.

- **grade**
  - Description: Overall grade of the house. Related to the construction and design of the house.
  - Type: Intial Object String (Transformed to Float) - 13 Vaues
  - Expectation/Comment: This paired with the condition may the the top useful in changing for prediction. Will be transformed into number from string later in model.

Modeling

Training and Testing & Cross Validation Approach

Predicting new home prices comes after training and testing the model. We will split our data set into 80% training set and 20% testing. The train/test split is used for intial model validation. Kfold cross validation will also be used. This allows the dataset to be split into "train" and "test" and then when the training data is used with cross validation it will be split into "training" and "validation" data.

The target value is the "price" value. This will be set to "y" and the rest of the data will be in "X". This is then used to do the initial train/test split.

We do not want the test data to be trained with the training data. This is data leakage and can distort the training process.

Baseline Model (1)

The initial linear regression model will be done with the highest correlated feature. This will be considered our baseline model. From here we will do several iterations to see if we can improve the model's performance with different techniques.

The highest correlated feature was the **"sqft_living"** feature.

The "sqft_living" feature has the highest correlation of .7 with the "price". The "grad_num" feature is the second highest correlated with the "price". This is good to know because this will be one of the attributes changed during predictions. The zipcode" feature has a negative correlation value. "Bathrooms" and "sqft_living are highly correlated as well. These correlations are the only ones above .7. This dataset doesn't have a really high correlation with the "price" feature or with each other.

Second Model with Categories and Numerical Features (2)

We will improve the baseline model by adding more features training more features(numerical and categorical). Additional features should support an increase a R2 score because it will help describe the dataset more. More processing of the data will occur across the training and testing data to support an improved model.

Feature Selection Modeling (3)

Certain features from the dataset will be transformed. Log transformations will be performed on the sqft features and the "price" to normalize their data, due to their skewness. Because we already transformed the "condition" and "grade" features to numbers, we will apply the One Hot Encoding transformation to the "Waterfront" categorical feature.

Home Value Distribution

Using the Stats model and Recursive Feature Evaluation (RFE) we will evaluate which features to eliminate.

Which features should be eliminated based on p value? We are assuming alpha level (significance level) of .05. If the p-value is above this, this let's us know we should reject these features. This is based on a hypothesis that these features are meaningful to the model.

Looks like there is a lot of multicolinearity in the model. The features that are above the needed p-value are:

- **"wa_NO"** .

Also, both "wa_NO" coefficient and the "zipcode" are outside of the confidence intervals, meaning they are outliers. The two most significant coefficients are the "grad_num" and "sqft_living based on their std_err and high t value

Ref Statsmodel InterpretationTim McAleer, 2020

Best Feature Selection/Final model (4)

The best features were used for the final model. After modeling three iterations at similar score of ~.65 for the training and validation data, a final model will be fit and used for prediciton using the last dataset.

The best fit model will be used to make predictions for homes that have home improvements. In this case the "condition" and "grade" features for a specific group of homes in certain zipcodes. The top five zipcodes that have the most homes with a **"condition" of 3-Good or less & a "grade" of 6-Low Average or less**". The chosen homes will all be modified to have a condition of "4-Good and a grade of 8-Good

Regression Results

Model Results

The goal was to create a multi-linear regression model that would be able to predict housing prices upon improvements. The model created to do this used an initial subset of data (from King County database) to process, train and test for this problem.

The final model was the fourth iteration. The intial model started as a base model linear regression model and increased in complexity with transformations (Log and One Hot Encoding) and filtering of features.

The final accuracy metrics are good enough to use the model for basic home value predictions.

RMSE Score Results:

The refinement of our model decreased the train and test RMSE scores by over 260K to ~.31. Applying the Log Transformer on our train and test target parameter "price" was the primary catalyst for this decrease. The price value was initial skewed and needed to be normalized.

The final model has a ~.31 RMSE. The closer RMSE to zero the better. The model increased its abilitity to accurately predict the target variable.

R2 Score Results:

The final model increased its R2 score by ~.15. The best base model score was .5 and the final test's model score ended with ~.66. As we refined the model, it increased its ability to account and explain for the variation. This was due primarily to the use of multiple features and the filtering of insignificant ones.

Linear Regression Model Assumptions:

The final model passed the linearity, normalization and homoescedasticity assumptions and failed the multicollinearity assumption.

For this particular problem we are using the model for predictions and inferencial uses. Therefore, the failure of multicollinearity was not a major roadblock this time. For inferencial use, a deeper evaluation on the most optimistic combinations of features to use for the model must be done.

Validation:

Our model development had a training, validation and test set. The test set that was intially split from the test set was segregated the entire model development. Using Kfold cross validation was a more robust validation method over the basic train/test split.

- Notebook Regression Results (Scores and Linear Assumptions)

Model RMSE and R2 Scores:

Final Train Mean Squared Error: 0.31441867034646886
Final Test Mean Squared Error: 0.3079357313675995

Final Train Model Score: 0.6438531698069864
Final Test Model Score: 0.6565129767577557

Third Train RMSE: 0.3147718460539656
Third Test RMSE: 0.3075890371545622

Third Train Model Mean Score: 0.644673314060476
Third Validation Model Mean Score: 0.6414594822403191

Second Model Kfold Train Mean score: 0.6446907194384797
Second Model Validation Mean score: 0.6414405396108452

Baseline Models:

Train RMSE: 260172.0361161922
Test RMSE: 268864.35998011974

Kfold Train Mean Score: 0.4884772214299433
Kfold Validation Mean Score: 0.5000072841051805

Train/Test Split Train Model Score: 0.49091149233831743
Train/Test Split Validation Model Score: 0.494749423259338

Prediction Results

Using the the trained multi-linear regression model, home prices were able to be predicted for a subset of homes. Because this problem was to predict home values after improvements, homes were chosen that had a major space for improvement.

The intial dataset used was filtered by the top five zipcodes that have the most homes with a "**condition" of 3-Good or less & a "grade" of 6-Low Average or less**".

The chosen homes were then modified to have a condition of **4-Good** and a grade of **8-Good** while all the other features stayed the same. The modified condition and grades were chosen as an objective goal. The thought was to give home owner incentives to improve upon their property, even if they do not reach the desired criteria.

These predictions will serve as samples for the real estate agents to understand the benefit of real estate data analysis. Because home value can depend on the surrounding community, grouping the homes for predictions by zipcodes helps the agent target a specific community for an extra benefit.

Home Improvement Prediction Results

Across the five chosen zipcodes, they all resulted in several hundred thousand dollar increase and minimum 100% increase in value.

Average home value differences :

- Zipcode 98118: $ 323,500, %100 Increase
- Zipcode 98106: $ 304,100, %100 Increase
- Zipcode 98126: $ 266,500, %100 Increase
- Zipcode 98146: $ 324,300, %200 Increase
- Zipcode 98168: $ 340,300, %200 Increase

Specific zipcode example (98188) had the highest amount of homes (106) that met our criteria for needing improvement.

Zipcode 98188 home value differences:

- Average of $ 323,000 price increase
- Lowest change of $ 14,500 increase (35,000 loss was recorded, but most likely an outlier error).
- Highest change $624,000 increase.

Average Zipcode Home Value Comparison

Average Zipcode Home Value Comparison

Conclusion

Limitations

Stakeholders Audience

There were several major limitations. A lot of it stemmed from understanding the domain and making processing and analysis decisions from that knowledge.

- **Limited dataset:**
  - The full dataset was filtered and scoped due to limited resources. There were attributes (features) in the dataset that would have supported a more accurate and model.
  - Dataset does not take into account aesthietics. There are no features or pictures evaluating this aspect.
- **Unknown realistic "Condition" and "Grade" values:**
  - Lack of knowledge on knowing whata realistic increase in "condition" and "grade" would be from the baseline. All chosen instances in the dataset were increased to the same values. It may be unrealistic to have a home in poor condition and below minimum building stards actually increase to "good" in both.
- **Unknown affects on other variables:**
  - The "condition" and "grade" feeatures wree the only ones modified from the intial dataset. Home improvements would definitely affect those features.

However, home improvements done on a home might also affect features like sqft_living, floors, bathrooms or bedrooms in some way.

- **Communal effects:**
  - The predictions were based on the modication of specific features for an individual home. The predicitons did not take into account how the home improvements would affect the surrounding comparible homes. This is common task done in real estate.
- **Model approach fit for specific problem:**
  - The model developed was for a specific problem. Understanding how home improvement how affect home values and the amount of the difference. Other usecases and problems were not taken into account. There may be insights to game from the model and predictions that
- **Time/Resources:**
  - Because of limited time and skillset a "good enough" model was delivered. The datasets, techniquesm model robustness had to be scoped.
- **Actual market culture:**
  - The dataset used wasn't the most current. These last two years have had major shifts in the supply and demand of homes. Therefore a shift in the home value estimates shift as well.
- **External effects:**
  - The model and predictions did not account for external effects that were not attributes of the home and its property.
- **Data scientist skillset:**
  - As a new data scientist the the model robustness is limited due to experience and skills.

Data Science Audience

- **Dataset limitations:**
  - A subset of the dataset was used. Using the whole dataset would have been would have given more choices to choose from in terms of correlation
  - The feature "waterfront" only had values of "Unknown", "No" and "Yes". It does not say what the waterfront view is, if the feature had a value of "Yes"
- **Linear Regression Assumptions:**
  - The multicollinearly assumption was the only assumption that clearly failed. Because of the resource and dataset scope limitations this aspect was not changed to fix this issue. If we wanted inferential analysis this would have to be fixed.
- **Limited in robustness:**

  - Not having the domian knowledge, a model that allowed for more flexibility in feature combinations and value valuations would have been beneficial.
- **Iterations:**
  - There were only 4 iterations done to create a prediction model. There were a lot of changes and approaches that would have required quite a lot more iterations to refine the model from the previous feedback received.
- **Feature combinations:**
  - The feature combinations were chosen from filtering the intial set down. Using the statsmodel p-value and the VIF. There wasn't a robust method that optimizes the best combinations for the model.
- **Model error and scores:**
  - It is not fully understood if the errors of a RMSE .3 R2 .66 are acceptable for the specific problem and solution approach. Also with these errors and score, it is still not understood what the cause of the of there errors is.

Recommendations

- Choose and present incinitives and vision of home improvement within a community. This incinitivizes people and helps with accountability.
  - Choose a few zipcodes to try out and then offer feedback that would improve model accuracy or approach.
  - Focus on major areas with high needs to bring up. The communal effect will possibly increase prices even though the highest improvements may not have happened.
  - Present businessess(e.i. construction, remodeling) of the potential work to be done. Partnering with them and possibly offering discounts to the select communities would be a good incintive for communities to improve together
- Market to protential homebuyers (individuals and investors) of the potential return on investment. These homebuyers may potentially buy the homes before the improvements and then fix them up.
- Increase consultation with the data scientiest/analys to improve our domain knowledge. As both parties educate each other the model solution has a better chance at being more accurate and robust.
  - Feedback on realistic feature values after home improvement
  - Having examples and case studies of home improvements specifics would help give a realistic picture to all of the stakheholder supporting the predictions and work.

Releases

No releases published

Packages

No packages published

Languages

- ● **Jupyter Notebook** 100.0%