

NAME OF THE PROJECT ------ Surprise Housing

Data Definition

A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them on at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia.

The company wants to know the following things about the prospective properties:

* Which variables are significant in predicting the price of a house
* How well those variables describe the price of a house.

Goal

Use the dataset after normalization, filling in missing values and adding new features with a linear regression model to predict sales prices.

## Method

There are a large number of features in this data set and in order to create a good prediction there will have to be a fair amount of work to do in advance. Combining the train and test set will means changes only need to be done once across both data sets. We will be exploring the following steps:

* Correlation
* Feature Exploration
* Missing Values
* Data Normalization
* Feature Engineering
* Assembling dataset
* Prediction.

There should be several good indicators for sale price, the challenge will be in normalization of the data and feature engineering

## 1. Data Understanding and Exploration

Let's first have a look at the dataset and understand the size, attribute names etc.

## The Data

The data has a wide range of integer, float and categorical information. In the end those will end up as numeric, but not quite yet. There are a couple of groupings of different types of measurements and understanding their differences will help in determining what new features can be created.

Neighborhood — Information about the neighborhood, zoning and lot. Examples: MSSubClass, LandContour, Neighborhood, BldgType

Dates — Time based data about when it was built, remodeled or sold. Example: YearBuilt, YearRemodAdd, GarageYrBlt, YrSold

Quality/Condition — There are categorical assessment of the various features of the houses, most likely from the property assessor. Example: PoolQC, SaleCondition,GarageQual, HeatingQC

Property Features — Categorical collection of additional features and attributes of the building Example: Foundation, Exterior1st, BsmtFinType1,Utilities

Square Footage — Area measurement of section of the building and features like porches and lot area(which is in acres) Example: TotalBsmtSF, GrLivArea, GarageArea, PoolArea, LotArea

Room/Feature Count — Quantitative counts of features (versus categorical) like rooms, prime candidate for feature engineering Example: FullBath, BedroomAbvGr, Fireplaces,GarageCars

Pricing — Monetary values, one of which is the sales price we are trying to determine Examples: SalePrice, MiscVal

## Correlation

A quick correlation check is the best way to the heart of the data set. There is a far amount of correlation for sales price with a couple of variables:

* OverallQual - 0.790982
* GrLivArea - 0.708624
* GarageCars - 0.640409
* GarageArea - 0.623431
* TotalBsmtSF - 0.613581
* 1stFlrSF - 0.605852
* FullBath - 0.560664
* TotRmsAbvGrd - 0.533723
* YearBuilt - 0.522897
* YearRemodAdd - 0.507101

## 2. Data Cleaning

Let's now conduct some data cleaning steps.

We've seen that there are some missing values in the dataset. We've also seen that variables are in the correct format, except some variables with distinct values, which should rather be categorical variables (so that dummy variable are created for the categories).

## Missing Values

There is a wide selection of missing values. First it make sense to hit the low hanging fruit first and deal with those that are missing a value or two, then work through what is left. Some of these look to be missing values not because they don’t have data but rather because the building was missing that feature, like a garage. Using Pandas Get.Dummies will sort that problem out into true/false values

## Replacing Missing Data

For the categorical information that is missing a single values a quick check shows which ones are dominant and manually replace the missing values. For most it is a quick process, I commented out the Python as I had in order to keep this shorter but feel free to uncomment and take a look.

### **Infer Missing Values**

Some of the missing values can be inferred from other values for that given property. The GarageYearBuilt would have to be at the earliest the year the house was built. Likewise TotalBasementSQFeet would have to be equal to the first floor square footage.

Once the missing values are filled in, it is good to confirm that the distribution did not go way out of wack. The blue is the original distribution, the green is the new one with missing values inferred and the red is the curve of the square root of lot area. I looks like a majority of these properties unsurprisingly do not have perfectly square lots but for our purposes, everything looks good.

## 3. Data Preparation

#### Data Preparation

Let's now prepare the data and build the model.

Feaure Engineering

Now it is time to create some new features and see how they can help the model’s accuracy. First are a couple of macro creations, like adding all the internal and external square footage together to get the total living space including both floors, garage and external spaces.

## Total Square Footage

There is a minimal increase of the Condition Rating to square foot, which is not a surprise. However when a linear regression with the Sale Price is taken instead a much more obvious pattern emerges.

## Total Rooms

Conversely if rooms were what interested you, there is a pronounced relationship between rooms and the number of bathrooms. This is the mean of bathrooms and the line is the variance (how much from the average most values deviate).

## Sale Month

Once last colorful graph before wrapping this up. I have wondered this about the Vermont housing market as the weather can have a large impact on many activities, including looking at houses I would imagine.

## Neighborhoods

All of us know that neighborhoods have a wide variety of characteristics. In some ways there is a relationship to sale price, but it is harder to define and less useful in prediction.

### **Features**

This is a way to edit what feeatures are included in the final model. I played a bit with which to include, and I have a feeling I may come back to this to re-engineer it.

### **Categorical Conversion**

In order for the model to understand categories, first replace all the categorical data with boolean values through Pandas get\_dummies.

Split Database

Time to split the database back into two parts, one with sales price and one without

## Training/Test Dataset

Create training set assembly

4. Model Building and Evaluation

Ridge and Lasso Regression

Let's now try predicting car prices, a dataset used in simple linear regression, to perform ridge and lasso regression.

## Model

Here are a variety of models you can try, many performed extremely poorly, Lasso was the best but I plan to go back to this and do further work on refining my features, like log regressions and normalizations.

## Conclusion

There were a couple of models tried. Random Forest Regression, LassoLarsCV and Ridge were all options and after experimenting I ended up settling on Ridge as the better option. What was fascinating and frustrating was the wide variation in accuracy each time I ran it. The models all seem to have some fluctuation in them. In the end I had about a 82%–90% accuracy. In trying to apply this same process to Burlington’s housing data I ran into even stranger inconsistency and huge over fitting problems. I have a feeling this will be a work in progress as I try it on other data sets.

Advanced Regression Assignment

Steps

1. Read the Data

2. See the info, stats to understand the data

3. Data Treatment

- removing columns with high no of missing values

- missing value imputation

- derived features creation (convert the year columns into age cols)

4. EDA - find the variables which are significant to the target variable

- pairplot - numerical vars with target variable

- heatmap - numerical variables - multicollinearity, target variable dependence

- goal --> analyze which variables are significant visually

5. Feature engineering

- label encoding

- one hot encoding

- target variable tarnsformation to normalize

6. Model Building

- Train-Test split (70%-30%)

- Scale data using Standard scaler

- Create X and y

- RFE [top 50 features]

- Build model for Lasso and Ridge

- Perform Gridsearch to get optimal values for lambda

- Re-run the model on test dataset using lambda optimal

Submitted by:

VENKATA NAGA SANDHYA NANDIRAJU

**ACKNOWLEDGMENT**

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

There were some great kernels out there that helped me work through the problem, definately check the following out: <https://www.kaggle.com/neviadomski/house-prices-advanced-regression-techniques/how-to-get-to-top-25-with-simple-model-sklearn>, <https://www.kaggle.com/poonaml/house-prices-advanced-regression-techniques/house-prices-data-exploration-and-visualisation> ,<https://www.kaggle.com/apapiu/house-prices-advanced-regression-techniques/regularized-linear-models> ,<https://www.kaggle.com/xchmiao/house-prices-advanced-regression-techniques/detailed-data-exploration-in-python>, https://www.kaggle.com/code/kefortney/house-prices-advanced-regression-techniques. https://www.kaggle.com/code/sid9300/assignment-surprise-housing-l-r ,https://www.kaggle.com/code/riteshpatil8998/top-49-house-pricing-using-ridge-and-lasso.

**INTRODUCTION**

* Business Problem Framing

Describe the business problem and how this problem can be related to the real world.

* Conceptual Background of the Domain Problem

Describe the domain related concepts that you think will be useful for better understanding of the project.

* Review of Literature

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

* Motivation for the Problem Undertaken

Describe your objective behind to make this project, this domain and what is the motivation behind.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Describe the mathematical, statistical and analytics modelling done during this project along with the proper justification.

* Data Sources and their formats

What are the data sources, their origins, their formats and other details that you find necessary? They can be described here. Provide a proper data description. You can also add a snapshot of the data.

* Data Preprocessing Done

What were the steps followed for the cleaning of the data? What were the assumptions done and what were the next actions steps over that?

* Data Inputs- Logic- Output Relationships

Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

* State the set of assumptions (if any) related to the problem under consideration

Here, you can describe any presumptions taken by you.

* Hardware and Software Requirements and Tools Used

Listing down the hardware and software requirements along with the tools, libraries and packages used. Describe all the software tools used along with a detailed description of tasks done with those tools.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Describe the approaches you followed, both statistical and analytical, for solving of this problem.

* Testing of Identified Approaches (Algorithms)

Listing down all the algorithms used for the training and testing.

* Run and Evaluate selected models

Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

* Key Metrics for success in solving problem under consideration

What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.

* Visualizations

Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

If different platforms were used, mention that as well.

* Interpretation of the Results

Give a summary of what results were interpreted from the visualizations, preprocessing and modelling.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Describe the key findings, inferences, observations from the whole problem.

* Learning Outcomes of the Study in respect of Data Science

List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how did you overcome that.

* Limitations of this work and Scope for Future Work

What are the limitations of this solution provided, the future scope? What all steps/techniques can be followed to further extend this study and improve the results.