HOUSING PRICES CASE STUDY

SUBMISSION

Avishek Sen Gupta

Lending Club

Introduction

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.

The company is looking at prospective properties to buy to enter the market. You are required to build a regression model using regularisation in order to predict the actual value of the prospective properties and decide whether to invest in them or not.

Business Understanding and Domain

The company wants to know:

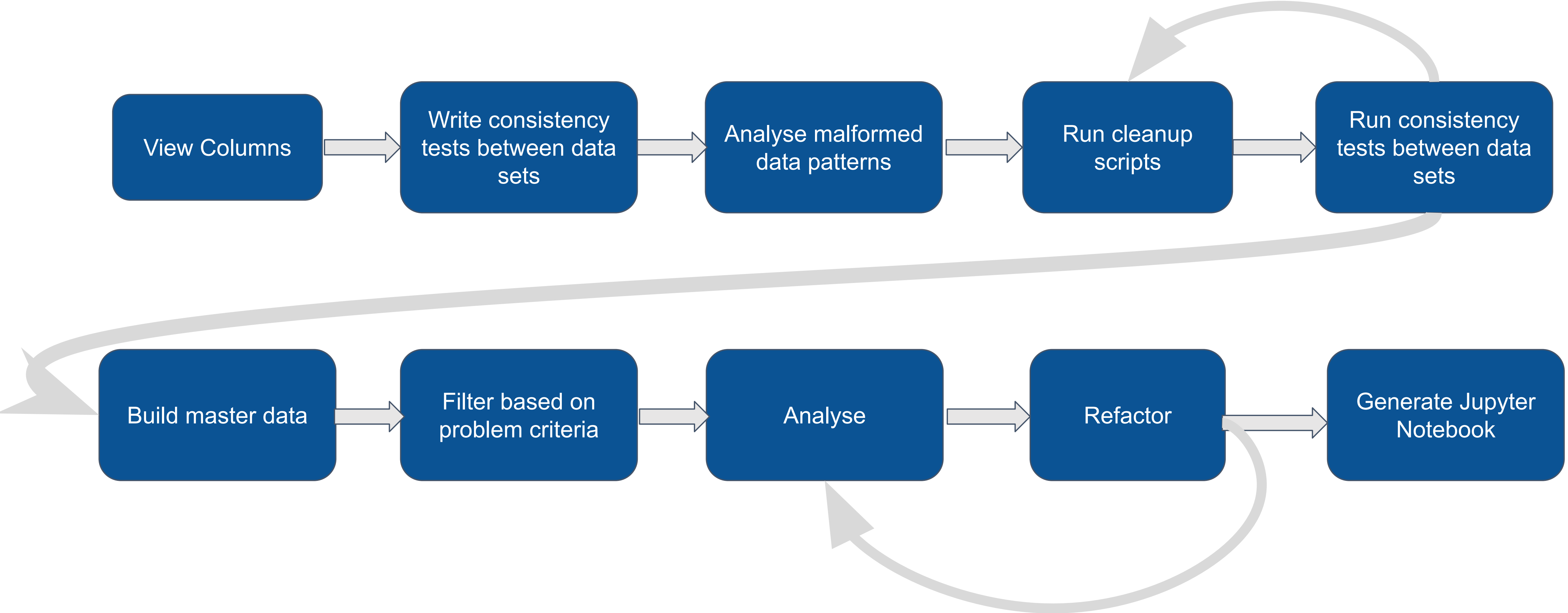
- Which variables are significant in predicting the price of a house

- How well those variables describe the price of a house.

Objective

We are required to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for management to understand the pricing dynamics of a new market.

Problem Solving Workflow



View Columns
Write consistency tests between data sets
Analyse malformed data patterns
Run cleanup scripts
Run consistency tests between data sets
Build master data
Filter based on problem criteria
Analyse
Refactor
Generate Jupyter Notebook

Data Understanding: CRISP-DM #2

Data understanding was achieved in several ways:

- A **data dictionary** was provided, which contained explanations for all the columns that are present in the original loans data set. This also directly helped in understanding what the valid values were in a particular columns, and later helped in the **Data Preparation** step.

The analysis is mostly focused on **Exploratory Data Analysis**, and some conclusions have been made, and some advice suggested. However, more rigorous **hypothesis testing** as well as potential **data transformations** (on data which could be normal but is skewed) **need to be made before drawing definitive conclusions** around **Driver Variables**.

Data Preparation: CRISP-DM #3

This is a summary of the steps we have taken to clean the data:

1. Dropping columns which are not useful for analysis, eg: Id
2. Checking for and imputing missing values
3. Fixing values which are inconsistent with what is presented in the data dictionary
4. Converting specific columns into dummy variables because they are categorical values
5. Converting specific column values into ordinal values because there is a valid ordering of values

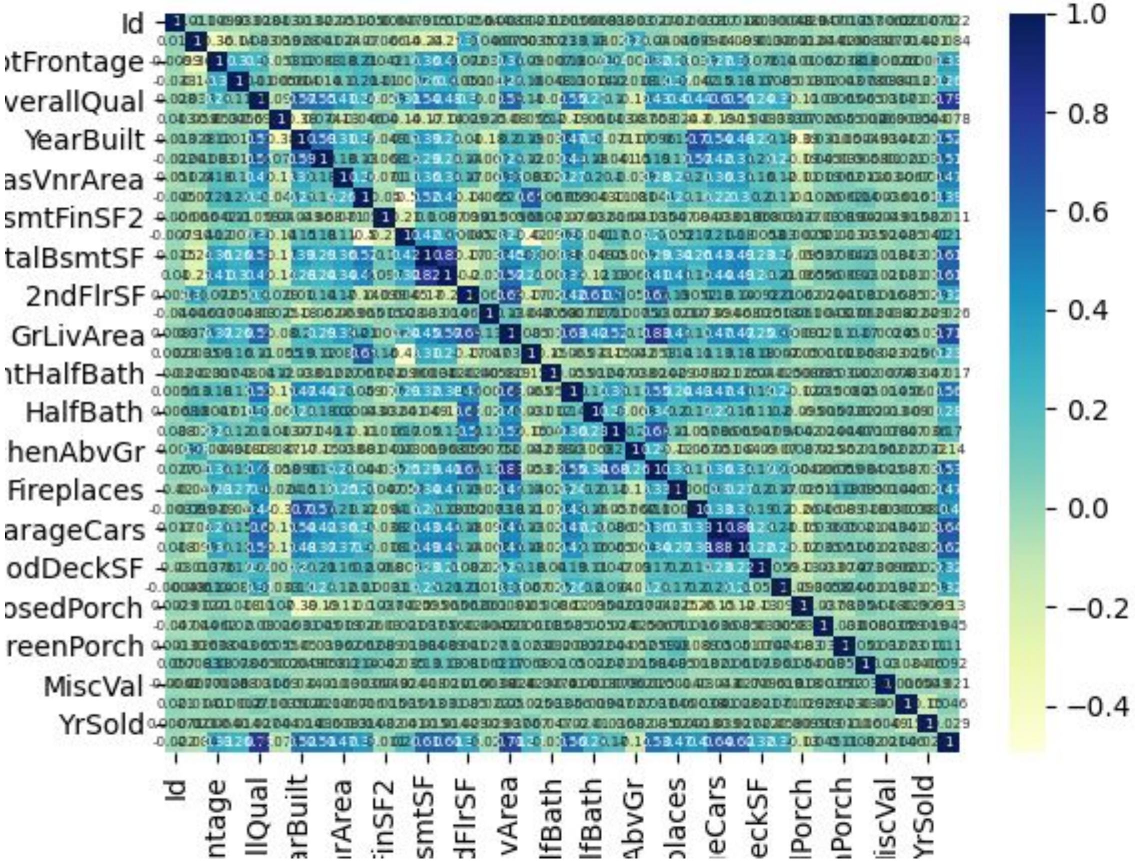
Data Analysis

Exploratory Data Analysis was divided into the following categories:

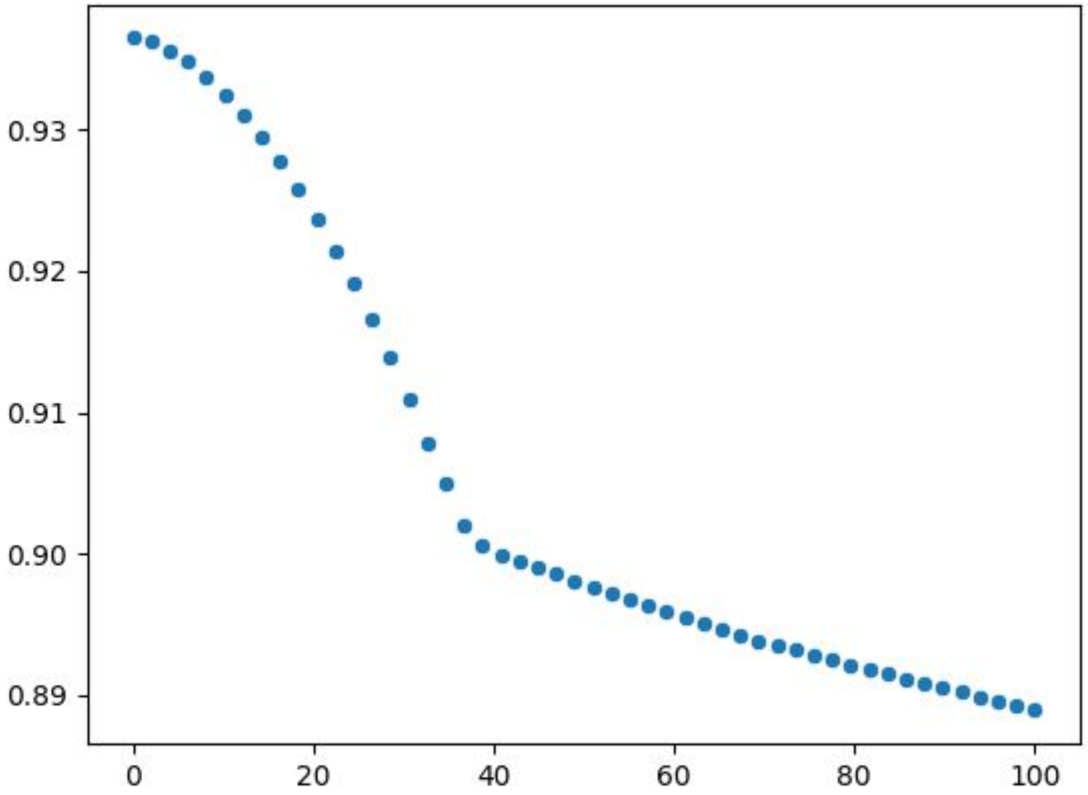
- Univariate Analysis: Checking **Distribution and Frequencies** of data
- Segmented Univariate Analysis: **Segmenting the data** as per loan_status column and performing univariate analysis
- Bivariate Analysis : Using **2 columns** at a time to see the relationship among the variables.

Findings

For Exploratory Data Analysis, a correlation plot was done across all the variables. No specific trend was visible.



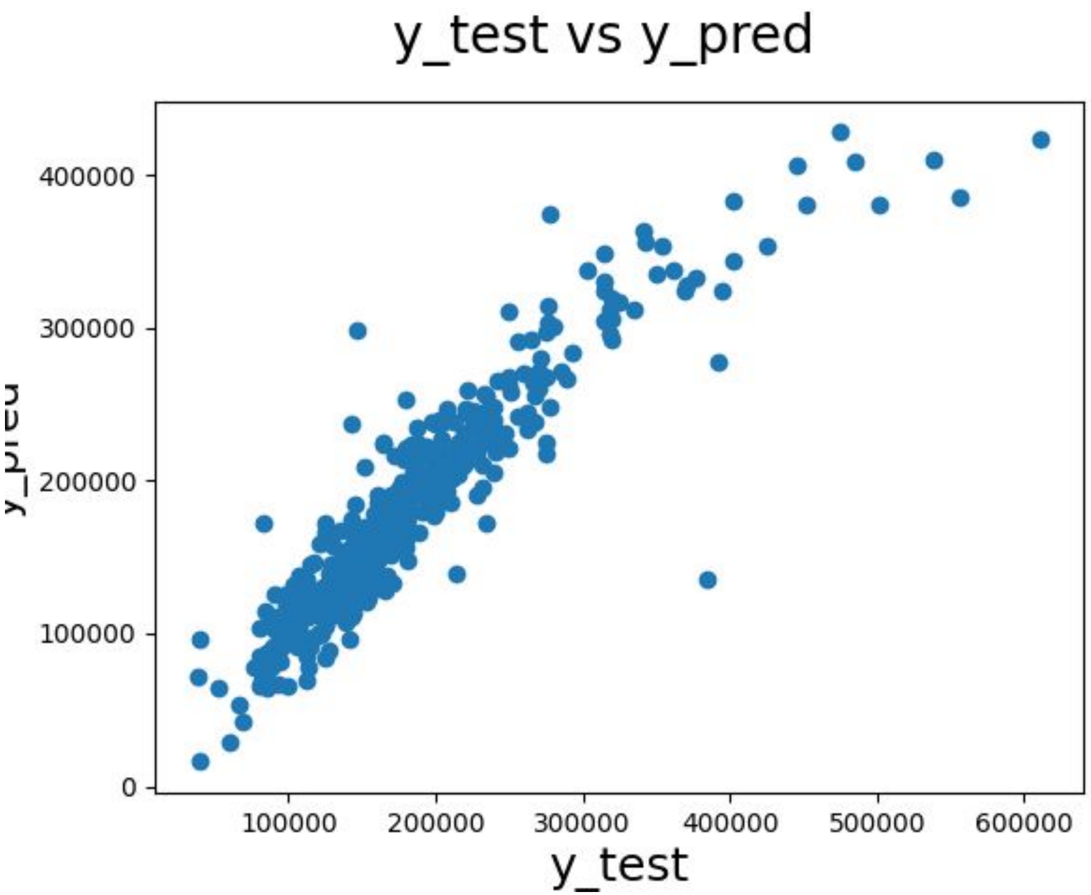
Findings



For LASSO Regression, the R2 value was plotted against the range of regularization coefficients. A value of coefficient around 40 gives an R2 score of 0.90, which is general enough to capture the macro trends.

The Ridge Regression technique showed no such trend.

Findings



For LASSO Regression, to test the goodness of prediction, the predicted y and the test y values were plotted on the test dataset.