HOUSING PROJECT

Submitted by:

Priyanka Priyadarshni Pradhan

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.

INTRODUCTION

- Business Problem Framing

  To build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not.

- Conceptual Background of the Domain Problem

  Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company. A US-based housing company named Surprise Housing has decided to enter the Australian market.

- Review of Literature

  Python Data Science Handbook Essential Tools for Working with Data Jake Vanderplas

- Motivation for the Problem Undertaken

  To build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Data Cleaning is done by preprocessing data checking missing values handling categorical data and quantitative data it is a regression problem and perform Exploratory Data Analysis (EDA)

  Correlations: It's often good to plot a correlation matrix to give an idea of relationships that exist in your data. It can also guide model building.

- Data Sources and their formats

  Data contains 1460 entries each having 81 variables. Two datasets are being provided to you (test.csv, train.csv). Train on train.csv dataset and predict on test.csv file.

- Data Preprocessing Done

  Duplicates & NaNs: I started by removing duplicates from the data, checked for missing or NaN (not a number) values. It's important to check for NaNs (and not just because it's socially moral) because these cause errors in the machine learning models.

  Categorical Features: There are a lot of categorical variables that are marked as N/A when a feature of the house is nonexistent. For example, when no alley is present. I identified all the cases where this was happening across the training and test data and replaced the N/As with something more descriptive. N/As can cause errors with machine learning later down the line so get rid of them.

  Date Features: For this exercise dates would be better used as categories and not integers. After all, it's not so much the magnitude that we care about but rather that the dates represent

different years. Solving this problem is simple, just convert the numeric dates to strings.

Decoded Variables: Some categorical variables had been number encoded.

- Data Inputs- Logic- Output Relationships

  Based on Input Model selection is done for prdiction

- Hardware and Software Requirements and Tools Used

  Python in Jupyter notebook

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  Supervised learning uses examples and labels to find patterns in data

  It’s easy to recognise the type of machine learning task in front of you from the data you have and your objective. We’ve been given housing data consisting of features and labels, and we’re tasked with predicting the labels for houses outside of our training data.

- Testing of Identified Approaches (Algorithms)

  Decision Tree — A tree algorithm used in machine learning to find patterns in data by learning decision rules.

  Random Forest — A type of bagging method that plays on ‘the wisdom of crowds’ effect. It uses multiple independent decision

trees in parallel to learn from data and aggregates their predictions for an outcome.

Gradient Boosting Machines — A type of boosting method that uses a combination of decision tree in series. Each tree is used to predict and correct the errors by the preceding tree additively.

Random forests and gradient boosting can turn individually weak decision trees into strong predictive models. They're great algorithms to use if you have small training data sets like the one we have.

Training

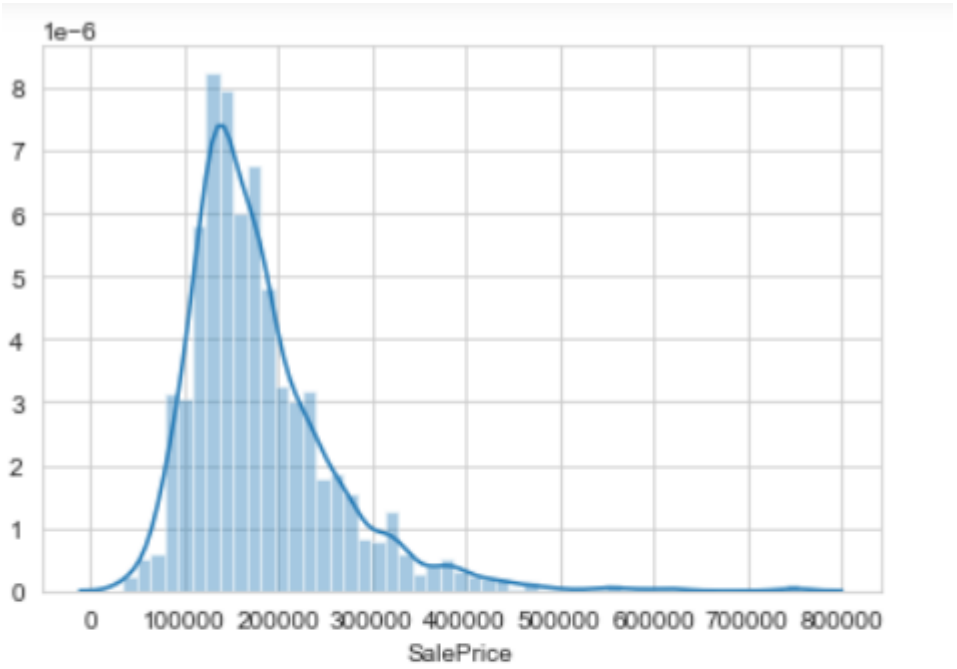
In machine learning training refers to the process of teaching your model using examples from your training data set. In the training stage, you'll tune your model hyperparameters.

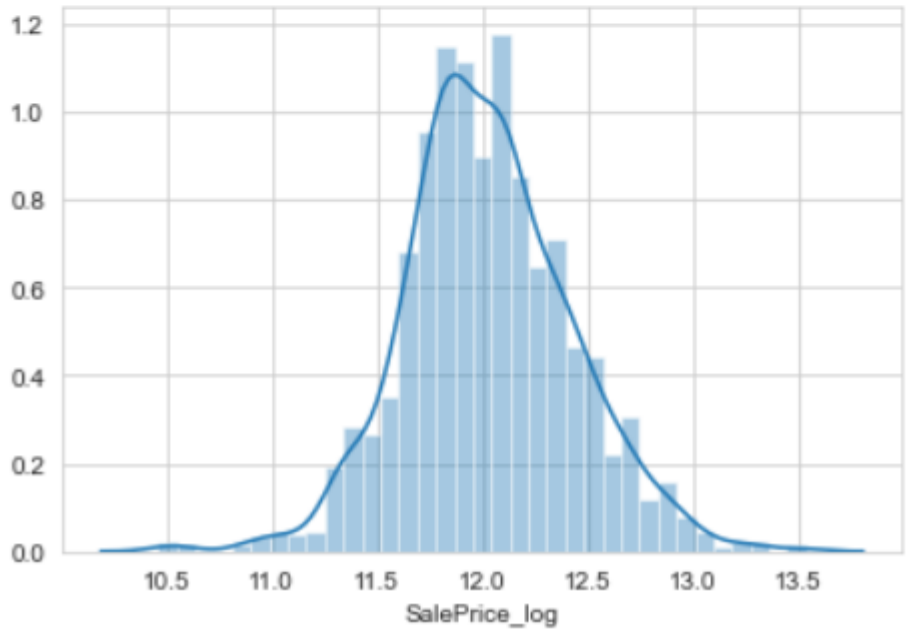
Before we get into further detail, I wish to briefly introduce the bias-variance trade-off.

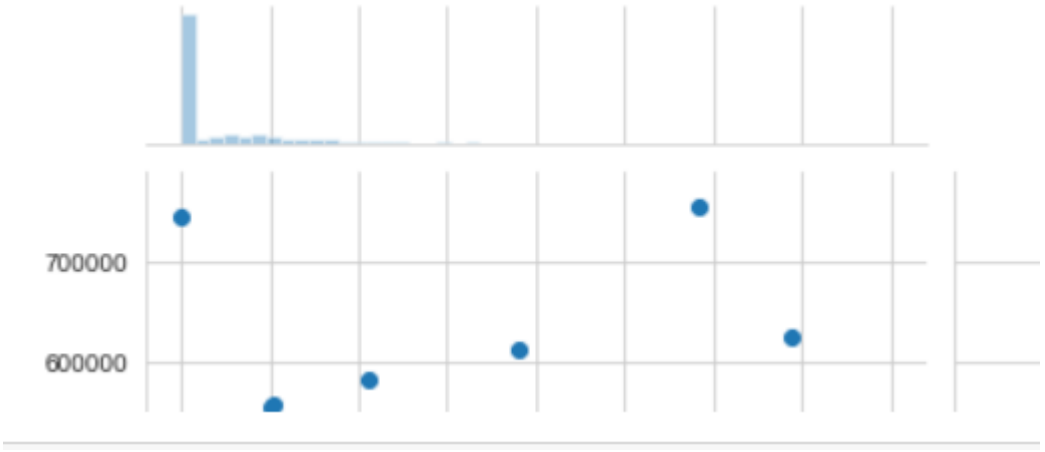
Model Bias — Models that underfit the training data leading to poor predictive capacity on unseen data. Generally, the simpler the model the higher the bias.

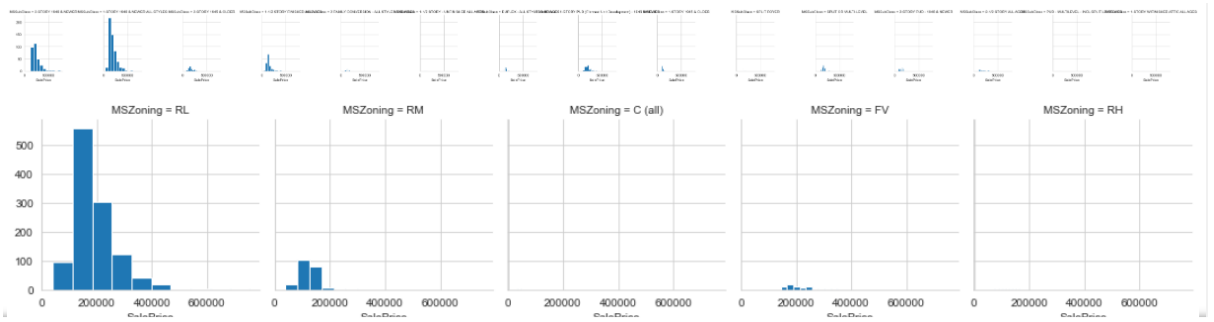
Model Variance — Models that overfit the training data leading to poor predictive capacity on unseen data. Generally, the more complexity in the model the higher the variance.

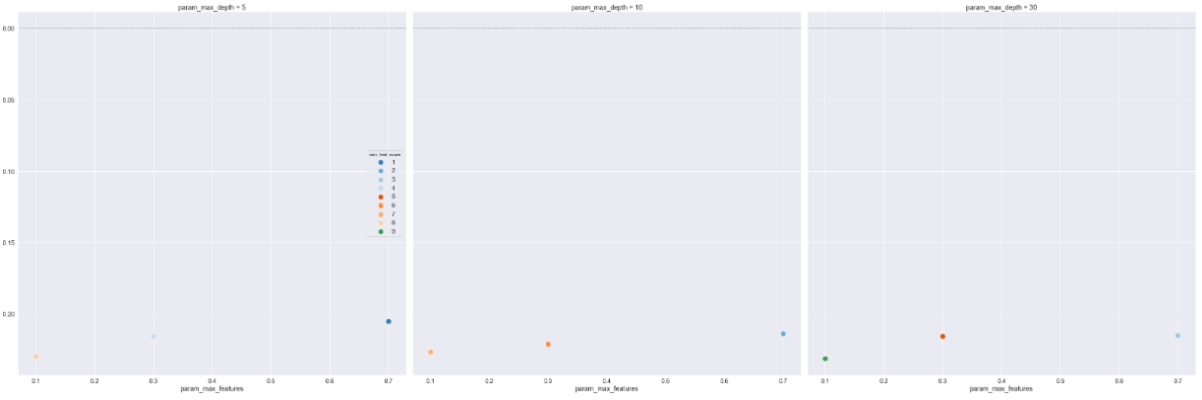
- Visualizations

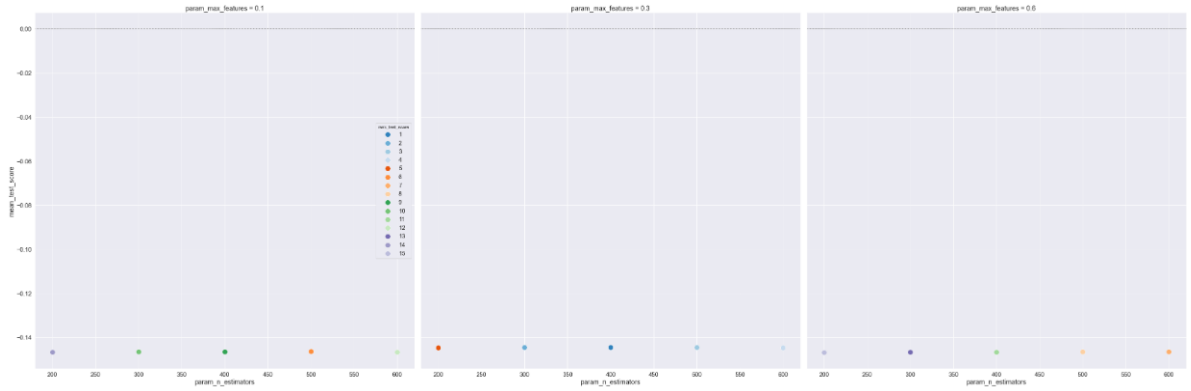

1e−6
8
7
6
5
4
3
2
1
0
0
100000
200000
300000
400000
500000
600000
700000
800000
SalePrice


1.2
1.0
0.8
0.6
0.4
0.2
0.0
10.5
11.0
11.5
12.0
12.5
13.0
13.5
SalePrice_log


700000
600000


MSZoning = RL
MSZoning = RM
MSZoning = C (all)
MSZoning = FV
MSZoning = RH


param_max_depth = 5
param_max_depth = 10
param_max_depth = 30
param_max_features


param_max_features = 0.1
param_max_features = 0.3
param_max_features = 0.6
param_n_estimators