FLIP ROBO

HOUSING PROJECT

Submitted by:

SAHIL SHAH

INTRODUCTION

- Business Problem Framing

  Housing prices are an important reflection of the economy, and housing price ranges are of great interest for both buyers and sellers. In this project, house prices will be predicted given explanatory variables that cover many aspects of residential houses. The goal of this project is to create a regression model that are able to accurately estimate the price of the house given the features.

- Conceptual Background of the Domain Problem

  The objective of the project is to perform data visualization techniques to understand the insight of the data. Machine learning often required to getting the understanding of the data and its insights. This project aims apply various Python tools to get a visual understanding of the data and clean it to make it ready to apply machine learning and deep learning models on it.

- Problem Statment

  Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world's economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. 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. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file below.

The company is looking at prospective properties to buy houses to enter the market. You are required 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. For this company wants to know:

- Which variables are important to predict the price of variable?
- How do these variables describe the price of the house?

Methodology

- Objectives

  1. Predict the sale price for each house.
  2. Minimize the difference between predicted and actual rating (RMSE/MSE)Data Sources and their formats

- About the data

  1. Number of data points in train data:1460
  2. Number of features in train data: 81
  3. Number of data points in test data: 1459
  4. Number of features in test data: 80

- Mapping the real world problem to a Machine Learning Problem

  This problem involves predicting the prices of the houses which are continuous and real valued outputs. Thus, this is a Regression Problem.

- Machine Learning Objective and Constraints

  1. Minimize RMSE.
  2. Try to provide some interpretability.

- Identification of possible problem-solving approaches (methods)

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

- Data Preprocessing

  Many real-world data-sets may contain missing values for various reasons. They are often encoded as NaNs, blanks or any other placeholders. Training a model with a data-set that has a lot of missing values can drastically impact the machine learning model's quality. Some algorithms such as scikit-learn estimators assume that all values are numerical and have and hold meaningful value. One way to handle this problem is to get rid of the observations that have missing data. However, you will risk losing data points with valuable information. A better strategy would be to impute the missing values.

- Handling Null Values

  To try and understand what the missing values are I looked at the data documentation this helped me transform other features to reflect the assumptions I made, for Example, GarageArea or Alley is zero, indicates we don't have a garage and cars should be transformed to 0 as well.

- Exploratory Data Analysis

  Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.We performed some bi-variate analysis on the data to get a better overview of the data and to find outliers in our data-set. Outliers can occur due to some kind of errors while collecting the data and need to be removed so that it don't affect the performance of our model.

LEARNING MODELS

- Logistic Regression

  Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. We can call a Logistic Regression a Linear Regression model, but the Logistic Regression uses a more complex cost function, this cost function can be defined as the 'Sigmoid function' or also known as the 'logistic function' instead of a linear function. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression.

- Linear SVM

  Support Vector Machine (SVM) is a supervised machine learning algorithm capable of performing classification, regression and even outlier detection. The linear SVM classifier works by drawing a straight line between two classes. All the data points that fall on one side of the line will be labeled as one class and all the points that fall on the other side will be labeled as the second. Sounds simple enough, but there's an infinite number of lines to choose from. How do we know which line will do the best job of classifying the data? This is where the LSVM algorithm comes in to play. The LSVM algorithm will select a line that not only separates the two classes but stays as far away from the closest samples as possible. In fact, the "support vector" in "support vector machine" refers to two position vectors drawn from the origin to the points which dictate the decision boundary.

- Decision Tree

  Decision tree is one of the predictive modelling approaches used in statistics, data mining and machine learning. A decision tree is a branched flowchart showing multiple pathways for potential decisions and outcomes. The tree starts with what is called a decision node, which signifies that a decision must be made. From the decision node, a branch is created for each of the alternative choices under consideration.

  Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks.

- Random forest

  Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

  As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

  The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

- Gradient Boost Classifier

  Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient boosted trees, which usually outperforms random forest. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

CONCLUSION

Final table if id and sales price has been successfully generated with the help of all four machine learning models.