Milestone 2

Group 8: Helen Zheng, Naman Agrawal, Sneha Thomas, Daniel Sampreeth Eadara.

Project Title: House Prices Prediction

Abstract

This is an open problem to predict housing prices that can be found on Kaggle. This is a classic problem to predict data based on a given dataset and can be extended to other problems. We hope to study how various methods such as linear regression, multivariate linear regression, random forest regressor, etc. can aid in predicting. In addition, we will attempt to incorporate additional information and use the models on different housing datasets.

Introduction

In this problem, which can be found on Kaggle as an open problem, the goal is to predict housing prices based on an existing dataset with attributes that describe the houses, such as garage size, neighborhood, condition, etc. Sales price is given as part of the training dataset, and the goal is to predict the sales price of the test dataset.

This is a classic problem to predict data based on a given dataset and features data to train a model as well as data to test on.

We are planning to apply various methods such as linear regression, multivariate linear regression, random forest regressor, etc. to see which of the methods gives a better result.

In addition, we will attempt to incorporate additional information, such as crime rate, distance to schools, and income level. We will also attempt to use the models on different housing datasets, such as in Boston and Manhattan.

Methods

Data

We found the coordinate data of schools and the houses and used a Google API to find the distance between each data. We incorporated this distance data to the Ames data set. We also plotted the sales price by month and tried to compare it to the average income in Iowa and the U.S. We also used yearly crime rate data of Ames.

Preprocessing

The dataset we used as input had a lot of null values, outliers and was showing some skewness. Since we want data that is as normally distributed as possible, we tried to reduce the skewness of those columns. Also all the null values and outliers were handled accordingly.

Feature Hacking/Selection

Since we have almost 80 features to base our house price prediction,we are using PCA(Principal Component Analysis) and LDA(Linear Discriminant Analysis) to reduce the curse of dimensionality and improve our model accuracy. We are also using correlation matrix to identify the relationship between different features and to identify the top features that contribute to sale price of the house. These three will help us in significantly improve our model accuracy by focusing more on the best features.

Prediction Model

Predicting house prices is a regression problem wherein we will have to predict the house prices,given a set of attributes like quality,year built,square feet etc. So we will be using the below regression algorithms to solve this problem:

1. **Multivariable Linear Regression**: This algorithm will help us in utilizing the linear relationship between the independent variables and dependent variable ‘SalePrice’. Even though it's a simple algorithm it can provide good accuracy rates. In addition, we can use regularization methods with this algorithm to prevent overfitting of data.
2. **Gradient boosting regression:** This algorithm will help in boosting weaker models by optimizing the loss functions.
3. **Random Forest Regressor :** RandomForestRegressor uses multiple decision trees and averages the value of these trees to generate a final output.Since this algorithm uses random sampling of data and random subfeatures to generate trees and their nodes, it is less prone to overfitting.
4. **Support vector regressor :** Support Vector Regression has a high capability of generalization and can result in high accuracy rates on unseen data.
5. **Ridge,Elastic Net,Lasso:** These algorithms are used to regularize the linear regression model and to improve the prediction accuracy of the model.

Model Training and Evaluation

We will be splitting the data in such a way that 80% data will be used for training and remaining 20% data will be used for testing. We will also be doing cross validation with 10 folds to make sure that the model will not show overfitting.

Accuracy scores like R2,RMSE will be used to evaluate the performance of the model.

Comparisons

Because this problem has been done several times, we plan to review the different methods and see which is better.

Conclusions

We plan to implement a solution to the housing price prediction problem found on Kaggle and review the different methods to see which is better. We also plan to incorporate the additional distance to schools data. We will then use the methods on different housing datasets, such as Boston and Manhattan.

Supplementary Information (Answers to your comments in milestone 1)

Various ways of approaching the problem have been done and are viewable on Kaggle. The methods used by others include Linear Regression, Random Forest Regressor, XGB Regressor, Ridge, and Lasso, to name just a few.

Our approach is different in the sense that we will be trying to compare the various methods used to see which is better, rather than simply aiming for the best result. We hope to study the various methods used in order to see which is better for this problem and see if we can rank them in order by most accurate.

We are working on adding to the dataset by including distance from schools, crime rate in the area, and average income level of those in the state and in the U.S.

The plan to clean the data include dropping columns more than a certain percentage of NAs and then dropping rows with NAs. For linear regression, we will filter the data that are non-numerical.

We will explore visualizing the data on a map per income range.

We have arbitrarily decided to split the data into either 50% training and 50% testing or 20% training and 80% testing.

We are looking at other datasets and have found some housing data in Manhattan, which has a drastically different range of sales prices.

So far, Sneha has done some preprocessing and has tried Linear Regression and RandomForest Regression to predict the house prices.She has also tried out PCA and LDA to see if dimensionality reduction will improve model accuracy. Daniel has taken a crack at pre-processing and looking at crime rate. Naman has found an external geocoded dataset of houses and has managed to find distance from coordinates of houses and schools nearby. And Helen has plotted income level by years corresponding to sales year as well as average sales price by month and year and has tried linear regression.

The remaining work to do include:

- Plotting sales price range onto the map to see if there is any correlation there.
- Implementing the various methods and comparing the results with each other.
- Implementing the various methods on the Manhattan data and comparing the results.
- Ranking the methods based on accuracy and concluding.

We are hoping to divide the remaining workload so that one person looks into plotting the sales price ranges onto a map and the remaining 3 people look for better ways to preprocess data and implement the various classification/regression techniques. Because there are many methods as described above, the remaining 3 people can each take on 2 methods and apply their pre-processing techniques.

In the conclusion, we will describe the methods and their pre-processing techniques and compare the results.