Diamonds Data Analysis and Prediction

Vishal Reddy Vennavaram S538349@NWMISSOURI.EDU

Northwest Missouri State University

Introduction

- Diamond's dataset contains various features line cut, clarity, carat etc.
- Using this dataset, I would like to predict the price of the diamond based on carat, clarity and cut.
- Cleaning and dimension reduction has been to remove the unused features.

Materials

Features:

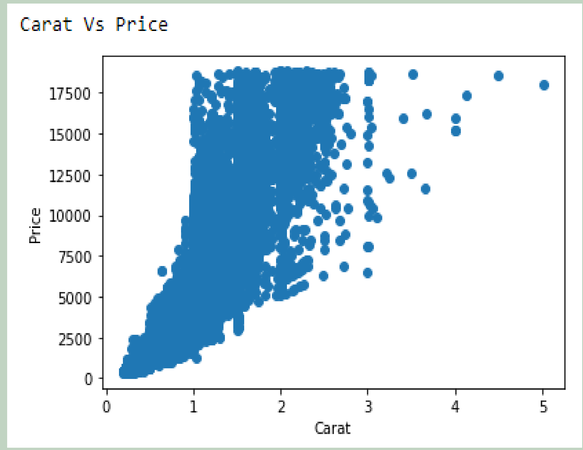
1. Data related to diamonds is used.

2. Information about diamonds like price, carat, clarity are used.

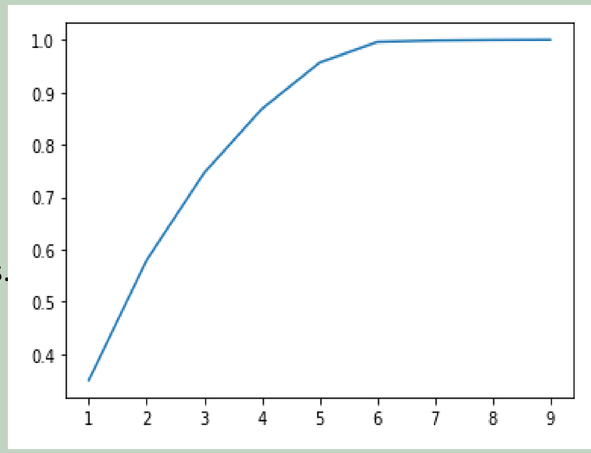
- Performed Linear Regression to train the model.
- Used SVM, decision tree for classification.

Results

The graph shown gives the Information about the relation between carat price.



The graph shown is a line graph between carat and price which shows that as the carat value increases the price of the diamond also increases.



Conclusion

- Found the relation between the price and various features like carat, clarity, cut.
- The Graphical representation shows that as the clarity of the diamond increases the price of the diamonds also increases.
- Accuracy of 99% has been obtained using a random forest classifier and the resultant f1_score value obtained is 0.99.
- The results are slightly better in the train set compared to that of test set which may be due to the overfitting of the data.
- Overfitting of the data can be avoided by using regularized data.

Challenges:

Performing analysis on continuous data is really difficult. So, I divided the data into categories by setting a range for the data.

Having string values in the dataset makes it difficult to perform various operations. So, I have converted them into integer values by setting values.

Additional Recourses

- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
- https://www.kaggle.com/shivam2503/diamonds
- https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

Acknowledgment

Mentor : Dr. Charles Hoot

Further Information

https://github.com/44-599-MachineLearning-S21/project-machine-learning-s21-Vishalreddy114