Computer Science Department

Final Project: Diamond Prices

Course: Introduction to AI (CS156)

Course Instructor: Yulia Newton, Ph.D.

Tanya Gupta

Introduction/Background

The dataset I will be working with has information about diamonds. The attributes of diamonds, like cut, clarity, and size will help us predict what the price of a certain diamond will be.

It is interesting because it is known that diamonds are one of the most valuable gems found in the world. They are not only used as jewelry, but they are used in windows, tools for cutting hard materials, medicines, audio equipment, and beauty products, contributing quite a lot to the economy.

Predicting prices based on the quality and the attributes of the diamond is important for diamond distributors and customers, which can help them budget.

This dataset will predict prices based on the physical attributes of the diamond, and does not have to do with the actual economy of the country (like the recession we have right now).

https://tech.hindustantimes.com/tech/news/diamonds-diamonds-everywhere-across-the-universe-yes-it-is-raining-diamonds-out-in-space-71662374036766.html

Dataset description

There are 9 independent variables.

Variable	Represents	Type of variable
carat	unit of weight	numeric
cut	stone shape	categorical
color	transparency grade	categorical
clarity	clarity scale based on impurities	categorical
depth	percentages of z/x	numeric
table	percentages of z/a, common parameter for diamonds, a not in dataset	numeric
x	X dimension in mm	numeric
y	Y dimension in mm	numeric
z	Z dimension in mm	numeric

The dependent variable is price. It is a numeric variable which represents USD. We will aim to train the data, and test to get the correct price. The price stands as a proxy for my prediction.

There are 43152 observations in training and 10788 observations in test in an 80-20 ratio. I also used cross-validation using 5 folds. My dataset has 53940 observations and will be regression, so I do not feel it's necessary to stratify.

Methodology

Algorithms

I decided to compare 3 algorithms for this project.

Random Forest Regression: Random Forest is a supervised learning algorithm that builds decision trees using a combination of learning models. This is often used to help things like customer activity and safety be predicted properly for industry efficiency. This is applicable here because diamonds and their prices affect customer activity.

Linear Regression: Linear Regression allows us to compare independent variables to a dependent variable, and can help us model our data to predict continuous data. This allows us to see the relationship between diamond features vs its price.

Linear Support Vector Regression: Linear Support Vector Regression is a type of SVM, a supervised learning algorithm, which maps data to feature space and categorizes data. It works well with data that has features X,Y,Z, which we have, so we will include this algorithm in our comparison.

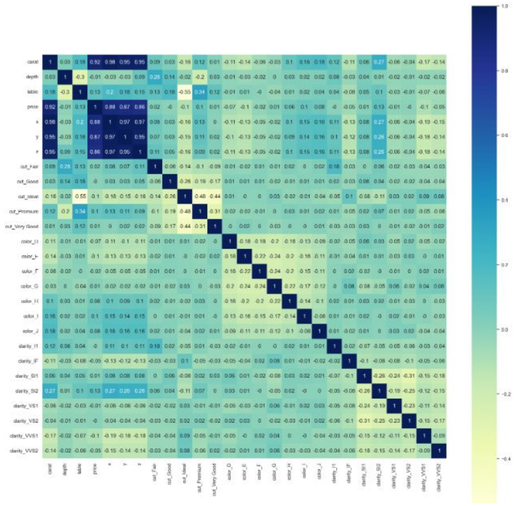

Figure 1: A heatmap to show the relationship between independent variables and dependent variables. We can see so far that size (x,y,z) seems to be closely related to price.

Application to the project

Since the independent variables consisted of both categorical and numeric variables, I used pandas.get_dummies to convert the features “cut,” “color,” and “clarity” into numeric data to apply to the regression models.

Analysis and Results

Pre-Analysis

For each independent numeric variable, I plotted a scatter plot with a best-fit line against the price. Features x, y, z, carat, and table show that as they increase, so does price. However, for depth it shows that as depth decreases, prices increases. For the categorical variables, I plotted barplots to see its relation with price. For example, the Premium category of cuts seems to have the most price total.

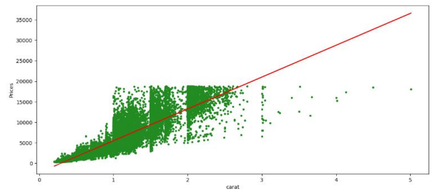

Figure 2: A scatterplot example of Carat vs Price.

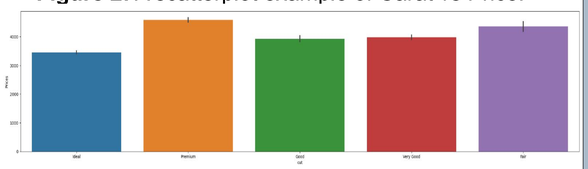

Figure 3: A barplot example of Cut vs Price.

Random Forest Regression Results

I cross-validated accuracy for the Random Forest Regression model. I also found the MSE, MAE, and R^2. Below are the values. The accuracy and coefficient are high, which is a good sign.

```
Mean Cross-Validation Accuracy for Random Forest Regression: 0.9786862170186899
Mean squared error for Random Forest Regression: 326791.60
Mean absolute error for Random Forest Regression: 280.89
Coefficient of determination Ytest vs Ypred for Random Forest Regression: 0.98
```

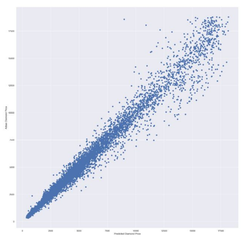

Figure 4: Random Forest Regression Scatter Plot (Actual Price vs. Predicted Price)

Linear Regression Results

I cross-validated accuracy for the Linear Regression model. I also found the MSE, MAE, and R^2. Below are the values. The accuracy and coefficient are not as high as Random Forest.

```
Mean Cross-Validation Accuracy: 0.9787474469758232
Mean squared error for Linear Regression: 1248486.71
Mean absolute error for Linear Regression: 737.44
Coefficient of determination Ytest vs Ypred for Linear Regression: 0.92
```

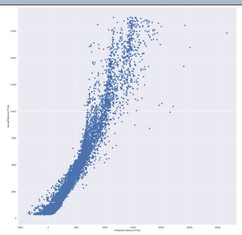

Figure 5: Linear Regression Scatter Plot (Actual Price vs. Predicted Price)

Linear SVR Results

I cross-validated accuracy for the Linear SVR model. I also found the MSE, MAE, and R^2. Below are the values. The accuracy and coefficient are the lowest compared to the other two algorithms.

```
Mean Cross-Validation Accuracy: 0.7915617441435849
Mean squared error for Linear SVC: 3096536.89
Mean absolute error for Linear SVC: 993.11
Coefficient of determination for Linear SVC: 0.80
```

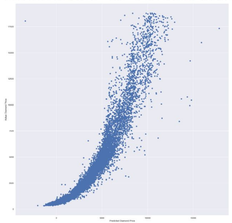

Figure 6: Linear SVR Scatter Plot (Actual Price vs. Predicted Price)

Summary/Conclusions

All three models perform pretty well, with Random Forest doing the best, and Linear SVC having the lowest coefficient and accuracy as expected. Random Forest has the lowest MSE and MAE out of the three algorithms, which also shows that the Random Forest model has performed the best.

In the future, I would like to experiment more with the Random Forest model, and see if selecting certain features instead of all features of the diamond would make prediction results better.

Key References

[1] Newton, Yulia. “Regression.Boston.ipynb.”

[2] Meltzer, Rachel. “What Is Random Forest?” *CareerFoundry.* https://careerfoundry.com/en/blog/data-analytics/what-is-random-forest/#:~:text=3.,%2C%20patient%20history%2C%20and%20safety.

[3] McGregor, Milecia. “SVM Machine Learning Tutorial.” *FreeCodeCamp.* https://www.freecodecamp.org/news/svm-machine-learning-tutorial-what-is-the-support-vector-machine-algorithm-explained-with-code-examples/

[4]Sharma, Gaurav. “5 Regression Algorithms You Should Know.” *Analytics Vidhya.* https://www.analyticsvidhya.com/blog/2021/05/5-regression-algorithms-you-should-know-introductory-guide/

[5]Giatti, Vittorio. “Diamond Prices.” *Kaggle.* https://www.kaggle.com/datasets/vittoriogiatti/diamondprices?resource=download

Acknowledgements

Thank you Professor Yulia Newton for a wonderful semester! I learned a lot about Artificial Intelligence, and that helped me with this project.