MMF1922 Data Science Report

Overview

The goal of the project is to predict prices of diamond based on the evaluation metric RMSE.

Data Summary

This dataset contains roughly 54,000 observations with the following variables:
Price: price in US dollars; Carat: weight of the diamond
Cut: quality of the cut with five categories (Fair, Good, Very Good, Premium, Ideal)
Color: diamond color with seven categories (D, E, F, G, H, I, J), note that D is the best and J is the worst
Clarity: a measurement of how clear the diamond is with eight categories (I1, SI2, SI1, VS2, VS1, VVS2, VVS1, IF), note that I1 is the worst and IF is the best; X: length in mm; Y: width in mm; Z: depth in mm; Depth: total depth percentage = $\frac{2*z}{x+y}$
Table: width of top of diamond relative to widest point

Exploratory Data Analysis

In this section, we want to give an initial assessment of our dataset, such as detecting data quality, checking statistical assumptions, and determining relationships among the explanatory variables.

From the first glance of dataset, we can see that there are lots of categorical variable that may need to be processed. We can also find that the deviation of price of diamonds is huge. The minimum price is 326, while the maximum price is 18823. Next, we want to check for missing value. We then calculate the percentage of missing values for each feature. Fortunately, our dataset is pretty clean without any missing values except for our response variable price. After that, we will check for our categorical variables to see if there's any category filled with relatively small counts. From the bar plots, we can see that all categories are meaningful, and we have no need to drop out any one of them.

Meanwhile, the distribution of our response variable price should be examined. From the histogram, we can see that the distribution is right skewed. Since the sample size is fairly large, it is still ok to conduct some statistical models and analysis under this circumstance. However, it is worth to try to transform the response variable. Finally, a heatmap is plotted to explore the relationships between price and other covariates. We can see that x, y, z, and carat are strongly correlated to price, while x, y, z, and carat are correlated to each other as well. This may impose a serious problem called multicollinearity. We will try ridge and lasso regression models to see what happens.

Feature Engineering

In this section, we want to do some feature engineering in order to make our modelling more smoothly and error-free. The main thing we do is to create new features that will help with modelling using one-hot encoding.

There are a total of three categorical variables in our dataset. The main characteristic of all of them is that they are ordinal variables with ordered categories. As a result, we need to assign an integer value to each unique category. For example, for variable cut, we will assign a "1" to "Fair", and a "5" to "Ideal".

Model Selection

In this section, we will try different models to see which one performs the best.

We start with the simplest model which is linear regression model. It performs not that well. Since we know that we have high multicollinearity, lasso regression is used to select a smaller set of features. Unfortunately, lasso has a bad result as well. We may need to try some other models. Then, linear support vector machine is fitted. It doesn't work, so there might exist some non-linear patterns of the dataset. This time, we will try some non-linear model to capture this pattern. A support vector machine with "RBF" and "Poly" kernels are built. Apparently, the SVM with "RBF" kernel works fine. In order to try more different models such as random forest, and ensemble methods, we use AutoML. The model selected by AutoML is Weighted Ensemble L2 with a very low RMSE.

Result

The final model we choose is Weighted Ensemble L2 with a RMSE of 507.

Model	Linear Regression	Lasso	Linear SVM	SVM "RBF"	SVM "Poly"	CatBoost	Ensemble L2
RMSE Train	1234	1234	1370	504	626	/	/
RMSE Test	1211	1210	1358	656	36862	519	507