Final Report

Introduction to Data Analytics

Project Title:

Prediction/Analysis of Price of Diamonds

Prepared By:

Simul Bista (N01489966)

ITE 5201 - Summer 2022
Humber College

Problem Statement

Prediction/Analysis of the price of diamonds

Dataset Description

- The dataset [1]contains the prices and other attributes of about 54,000 diamonds. The information is based on the 2017 pricelist from Tiffany & Co's[2] – an American luxury jewelry retailer.
- The following dependent variables were trained to predict/analyze the price of the diamonds($USD).
  - carat weight of the diamond (0.2--5.01)
  - x length in mm (0--10.74)
  - y width in mm (0--58.9)
  - z depth in mm (0--31.8)

Dataset Analysis & Observations

- Initially the data was cleaned (data with faulty dimension values were removed and then outliers were also taken care of)
- Univariate and Bivariate analysis were carried out to understand the correlation between the features and the to-be-predicted variable i.e., the diamond price.
- Observations

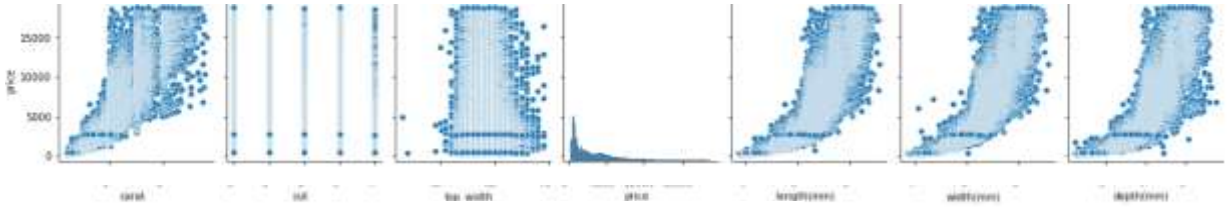

Fig 1: Pair plot between prices and other features

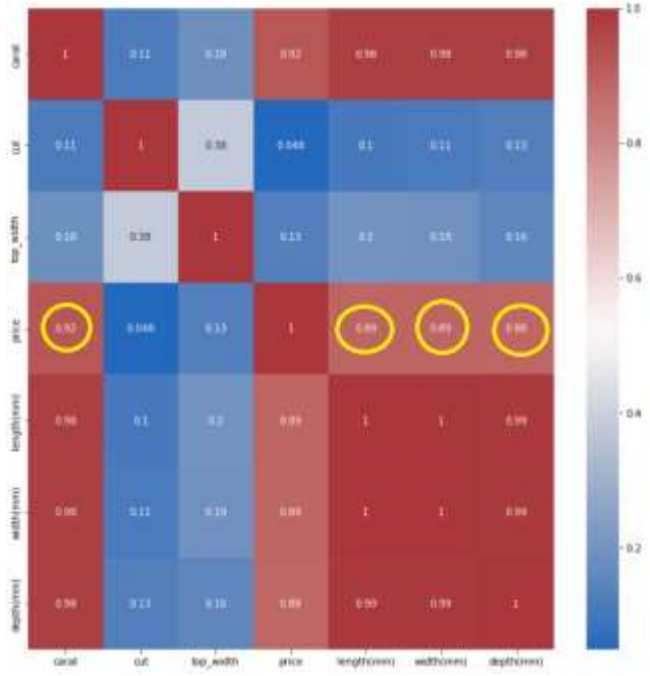

Fig 2: Heatmap (correlation) between the features

- The pair plot (Fig.1) shows that price has **strong, non-linear, and positive correlation** with the carat and dimension features of the diamond.
- The heat map (Fig. 2) bolsters the previous claim of correlation (correlation coefficient of **0.92** between prices and carat, and **0.89** between prices and dimensions).
- Finally, the univariate analysis of price (Fig. 3) shows **positive skewness** i.e., the frequency of the price starts to decrease as it gets higher.)

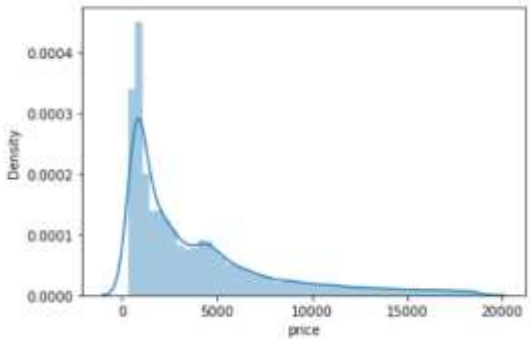

Fig 3: Univariate Analysis (price)

[1] https://www.kaggle.com/datasets/shivam2503/diamonds| Kaggle. Accessed July 30,2022
[2] Tiffany & Co. Official | Luxury Jewelry, Gifts & Accessories Since 1837. Accessed July 31,2022

Proposed Analytical/Prediction Model

- There were 4 dependent and continuous variables that showed correlation to the price of the diamonds, hence Linear and Polynomial Regression Models were preferred.
- For Polynomial Regression Model, degree 1 was used first and then degree 3 was used in the second iteration which swayed away from overfitting behavior and hence better result (with less errors) was achieved.

Results & Discussions

- Fig. 4 shows linear regression model plot whereas Fig. 6 and 8 show polynomial regression models plots with degree 1 and 3, respectively.
- The **polynomial model (with degree 3)** performs the best with the **least errors** (with degree 1 being underfitting, and degree 2,4,5 and above being overfitting).
- The comparison between the errors from the two regression models can also be seen.

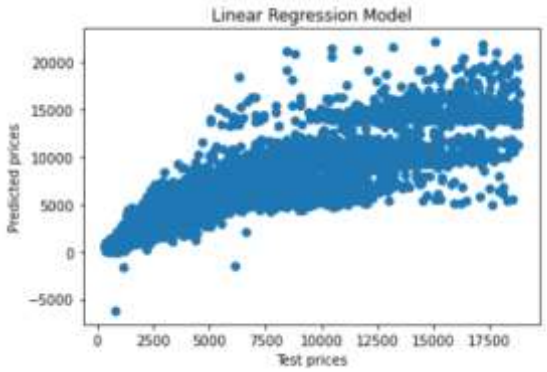

Fig 4: Linear Regression Model

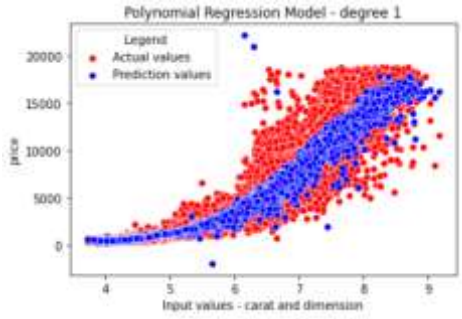

Fig 5: Polynomial Regression Model (degree=1)

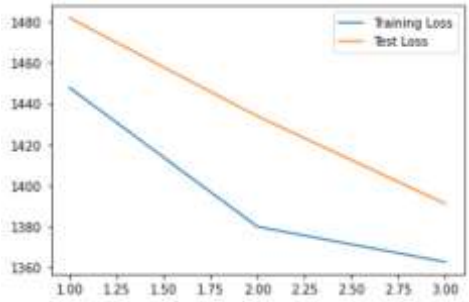

Fig 6: Determining the best value of degree

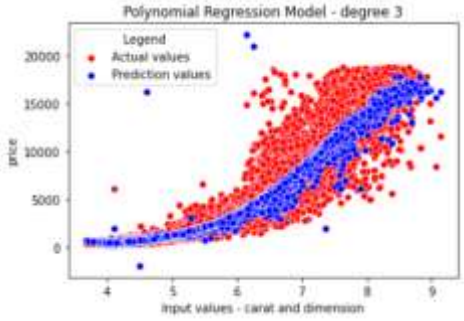

Fig 7: Polynomial Regression Model (degree=3)

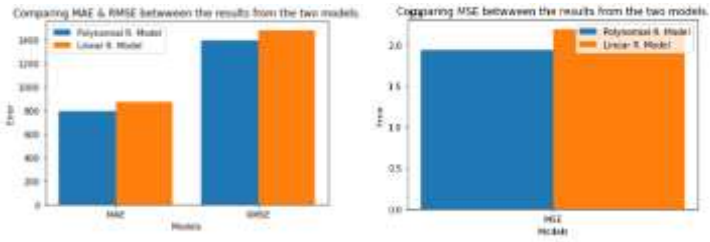

Fig 8: MAE, RMSE and MAE comparison between linear and polynomial regression models

Conclusion

The above analysis concludes that features such as **carat size** and the **dimension** impact the price of the diamonds whereas features such as cut and top width does not. Moreover, **improved data accuracy** was observed in the **polynomial regression model (of degree 3)** in comparison with linear model and polynomial model with degree other than 3.