Rupal Gandhre

Data Science Capstone2 Final Report

Background:

The jewelry industry has a potential to benefit from data and advanced analytics. Many of the retail industry sectors are already leveraging the benefits. With the current COVID-19 impact, most of the sales have been through ecommerce websites.

Goal:

The goal of this project is to classify the price range for jewelry based on the features of jewelry. The features include

1. Metal of jewelry (18K Gold, 14K Gold, Sterling Silver)

2. Type of Stone(Diamond or Gemstones)

3. Color of the Stone

4. Cut of the Stone

5. Carat weight of the Stone

This model may help the client to get a price range for the custom jewelry

Criteria for success: Built a classification model for the price and complete the capstone by November 2021

Scope of solution: Find the correlation between Price and Metal, Stone, Stone Color, Stone Cut, Stone Clarity for Rings jewelry type

Constraints: Lots of inconsistent and missing data.

Real time data for Sales, Inventory, Operation Cost is not available .

Stakeholders: Rupal Dukhande, Dr. Ruturaj Soman

Data sources:

1. Web scraping https://www.effyjewelry.com/collections/all using BeautifulSoup

Dataset:

I did not find datasets related to the jewelry industry, so I decided to learn how to web-scrape data using BeautifulSoup. The data was scraped from https://www.effyjewelry.com/collections/all

I am thankful to the web developers for not implementing a script to block my nuisance of an IP address.

SourceCode for scraping:

https://github.com/rgandhre/SpringBoard/blob/main/DataScience_Capstone2/notebooks/Scrapping_Effy_Pages.ipynb

https://github.com/rgandhre/SpringBoard/blob/main/DataScience_Capstone2/notebooks/Scrapping_Effy_Pages.ipynb

Data Wrangling:

https://github.com/rgandhre/SpringBoard/blob/main/DataScience_Capstone2/notebooks/Effy_Data_Cleaning.ipynb

Output File:
https://docs.google.com/spreadsheets/d/15oVVmmskmgShgglffbH5On8b_8jPv3Xno3f6igwOQmc/edit?usp=sharing

Following fields were scrapped at first:

Column_Name	Description	Action Taken
Description	Description of the jewelry	Dropped later on
Jewelry_Type	Type of jewelry(Earring, Rings, Bracelet)	Data split based on the jewelry type
Discount_Price	Sale price of the jewelry	Dropped later on
Price	Actual price of the jewelry	This is the target variable. It is categorized in range 'under_2000', 'above_2000_and_under_3000', 'above_3000_and_under_4000', 'above_4000_and_under_5000', 'above_5000_and_under_6000', 'above_6000_and_under_8000', 'above_8000'

Metal	Metal used to make the jewelry (example 14K Gold, Silver etc)	
Metal_Color	Color of the metal(Rose, Yellow, White)	Dropped later on
Stones	Details of the stones used in the jewelry Stone name, Stone Color, Stone Cut, Stone Carat Weight	This field was split to make more features for each stone(n) Stone[n]_Stone Stone[n]_Color Stone[n]_Cut Stone[n]_Carat This are the primary features used to predict the price range of the jewelry

After doing distribution analysis, the majority of the data was for Jewelry_type = Rings. Me and my mentor(Raja) decided to do exploratory data analysis for Rings jewelry type.

Exploratory Data Analysis:

Source Code:

https://github.com/rgandhre/SpringBoard/blob/main/DataScience_Capstone2/notebooks/Effy_EDA_Rings.ipynb

Input File:

https://docs.google.com/spreadsheets/d/15oVVmmskmgShgqIffbH5On8b_8jPv3Xno3f6iqwOQmc/edit?usp=sharing

1. Dealing with null values:

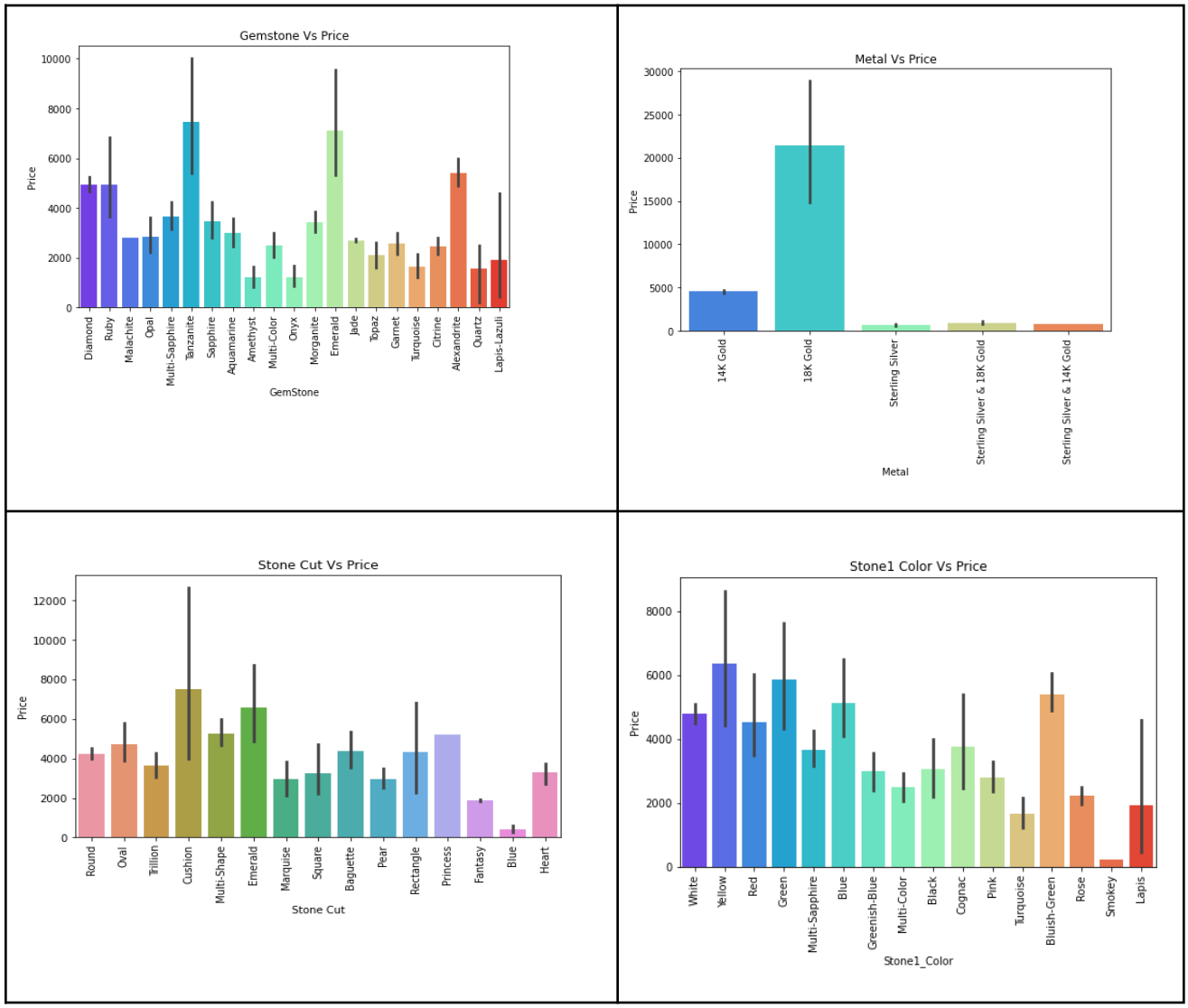
   From the rings dataset, the majority of the rings had 2 stones or data for more than 3 stones was not available. These records were deleted. For null values in Stone Color, I used google search to find the natural color of the stone and updated the values accordingly.

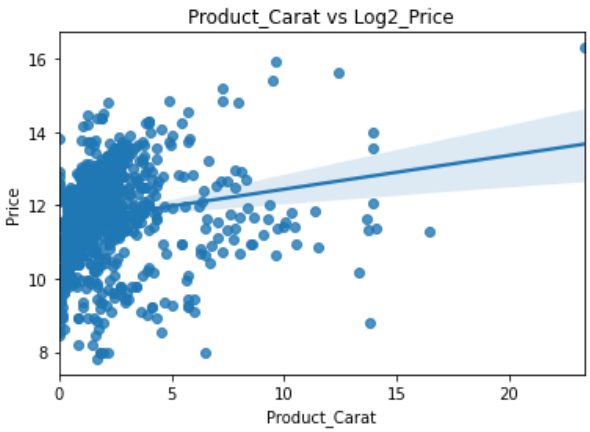
   At this stage, data was split into following columns

Column_name	Action take if null

Jewelry_Type	Only records with Jewelry_Type='Ring' was selected
Metal	Null records were deleted
Stone1_Stone	Null records were deleted
Stone1_Cut	Null records were updated with default value as 'Round', as majority of the data was 'Round' cut
Stone1_Color	Null records were updated with the most commonly color the stone is found as per google research
Stone1_Carat	Null records were deleted

2. Exploring Data Correlation





Pre-processing and Training:

Source Code For Diamond Dataset:

https://github.com/rgandhre/SpringBoard/blob/main/DataScience_Capstone2/notebooks/Diamond_Rings_preprocessing_training.ipynb

Input File:

https://docs.google.com/spreadsheets/d/1LoY3hOoBxGRZ9iDaVdYkfTa3-xKu1Cgic_B9mkhPG_4/edit?usp=sharing

Source Code for Gemstones dataset:

https://github.com/rgandhre/SpringBoard/blob/main/DataScience_Capstone2/notebooks/Gemstones_Rings_preprocessing_training.ipynb

Input File:

https://docs.google.com/spreadsheets/d/1HhwsJyNRhgXqYJZMIGlf9rBW3E4q82V5YI21m8CEtSo/edit?usp=sharing

50% of Rings were made of ‘Diamonds’ . So the data was again split between ‘Diamonds’ and other ‘Gemstones’.

Since this is a classification problem, following classification models are used and compared for their precision/recall score.

1. Logistic Regression

2. Random Forest

3. Stochastic Gradient Descent

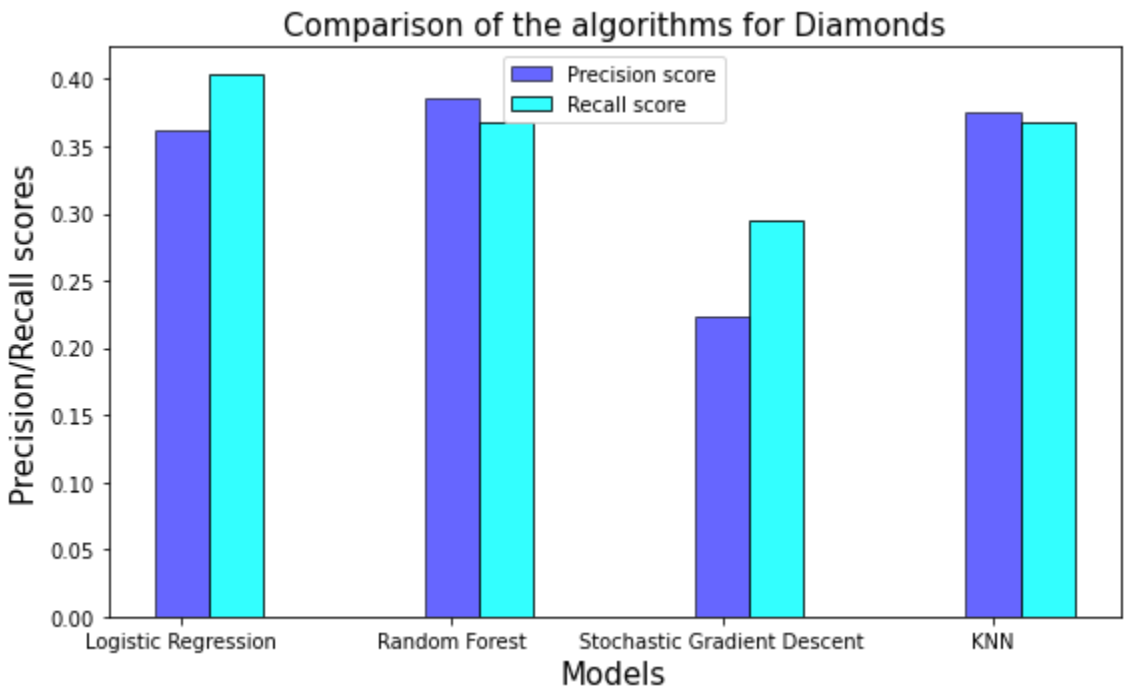
4. K-Nearest Neighbor(KNN)

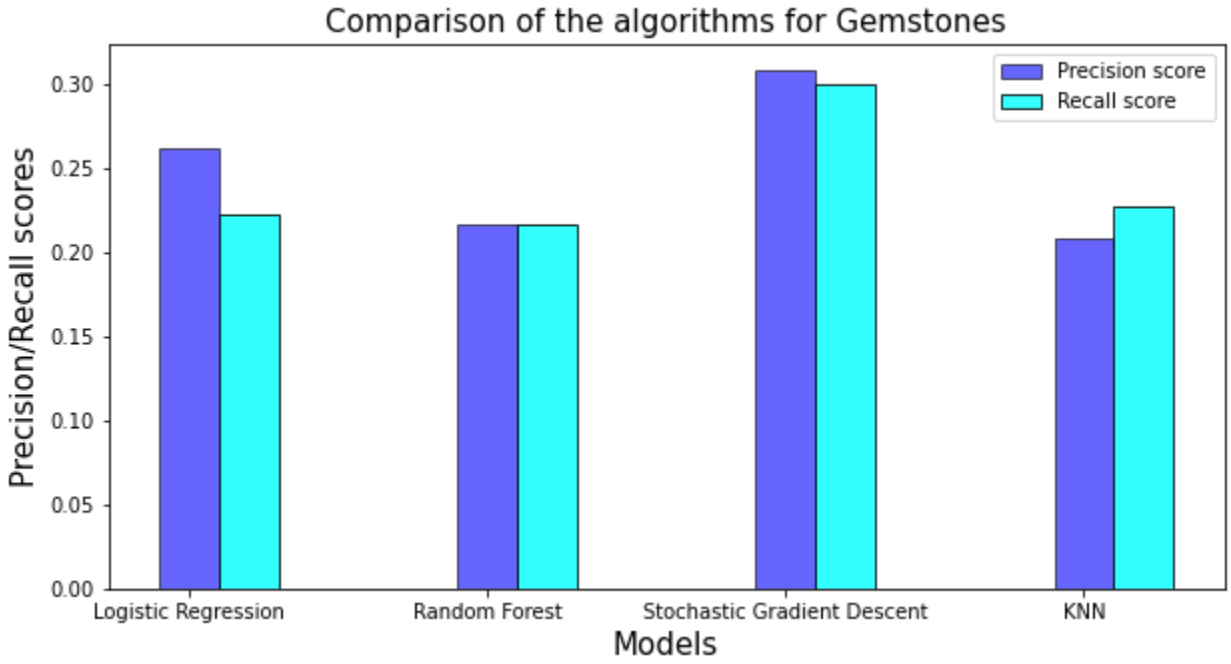
Evaluating the performance of a model by training and testing on the same dataset can lead to the overfitting. Hence the model evaluation is based on splitting the dataset into train and validation set. But the performance of the prediction result depends upon the random choice of the pair of (train,validation) set. Inorder to overcome that, the Cross-Validation procedure is used where under the k-fold CV approach, the training set is split into k smaller sets, where a model is trained using k-1 of the folds as training data and the model is validated on the remaining part.

All of the columns of the data are categorical. One-hot encoding is used to convert categorical data to numeric before training the data. .

Since the dataset is not too large, GridSearchCV is used for hyperparameter tuning. GridSearchCV is a library function that is a member of sklearn's model_selection package. It helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So, in the end, you can select the best parameters from the listed hyperparameters.

Evaluation of the Model is done using Confusion Matrix and Classification Report


Comparison of the algorithms for Diamonds
Precision score
Recall score
Precision/Recall scores
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Logistic Regression
Random Forest
Stochastic Gradient Descent
KNN
Models


Comparison of the algorithms for Gemstones
Precision score
Recall score
Precision/Recall scores
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Logistic Regression
Random Forest
Stochastic Gradient Descent
KNN
Models

Conclusion:

The classification models were evaluated based on the precision and recall score.

With the currently available dataset,

1. For diamond rings models, the precision score for Logistic Regression is 0.361152 and thus is selected as the best model
2. For gemstones rings models, the precision score for Stochastic Gradient Descent is 0.308762 and thus is selected as the best model.

With more diverse and accurate data, the model may predict better results.

Future Work:

The model has a lot of room for improvement.

1. Get more accurate and diverse data for all jewelry types
2. In-depth analysis and modeling was done only for rings.
3. Data for other jewelry types needs to be analyzed, explored and modelled.
4. Due to limited data available on the product features, predictive sales price could not be done. If available, analyze more financial data to predict the sales of the company.