The University of Memphis
Computer Science
COMP 7150-Fundamentals of Data Science
Project Proposal: **African Diamond Price Prediction in Memphis**
Rahul Marru, Data Science, Fall 2022

PROJECT GOAL AND SCOPE

The goal of the project was to solve the problem of predicting the market price of African diamonds in Memphis a month in advance**.** The solution was proposed to be a working python program with a relative error (RE) of at most 20% on average.

BACKGROUND/LIT REVIEW

The authors, Abirami R and Agniswar P (2021) [3] made a study that Diamonds are one of the most valuable gems in the world. It is also one of the most expensive gems, and therefore has an extremely volatile price. The value of diamonds depends upon their structure, cut, inclusions (impurities), carats, and many other features. Diamond prices are usually set for the day and traded in US Dollars. To better predict the price of diamonds, the Kaggle diamond dataset is used and a scatterplot of metrics such as carats, price, and the color is used to understand the nature of their relationships. They solve the problem of forecasting diamond prices in the United States. They use four prediction algorithms based on this understanding: linear regression, lasso regression, support vector machines, and random forest. From that Random Forest has the most accurate results since it makes use of multiple trees, with 98% accuracy. Secondly, lasso regression has an accurate result almost equal to linear regression. Support Vector Regression (SVR) has the worst result of 68% accuracy. The shows the r2 score is 90% which means the score is good and the prediction is very accurate. The mean Square error is 0.151 which means that we have a very small margin of error since 0 is no error and1 is a high margin of error. Mean Absolute Error is around $805 which is a very small error since the values of diamonds are in hundreds and thousands of dollars. Explained Variance Score is 0.905 which is almost 1, which is the best score.

Soham Jani and Ruchi Gajjar (2021) [4] authors have researched that Diamond is one of the strongest and the most valuable substances produced naturally as a form of carbon. However, unlike gold and silver, determining the price of a diamond is very complex because many features are to be considered for its price. Both made a study on the prediction of the price of diamonds in the United States. How machine learning can predict the price of the diamond you desire to buy and the basic measurement metrics to measure the quality of prediction algorithms. The comparative analysis of various machine learning Regression models - Linear regression, Support Vector regression, Decision trees, Random Forest regression, K-Neighbors regression, CatBoost regression are done for the price prediction of any diamond. From the performance parameter values and analysis, it was found that the CatBoost Regression algorithm proved to be the most optimal algorithm having an R2 score of 0.9872 and formidable training and testing accuracies of 98.74% and 98.72% respectively.

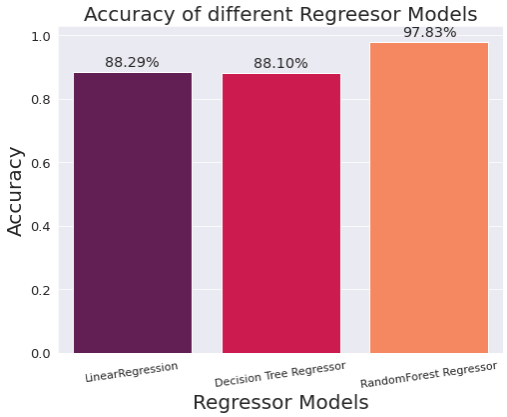
According to articles [1] [2], Gemstones like diamonds are always in demand because of their value in the investment market. This makes it very important for diamond dealers to predict its accurate price. However, the prediction process is difficult due to the wide variation in the diamond stones sizes and characteristics. Where the authors solved the problem of predicting diamond prices in the United States by several machine learning algorithms such as linear regression, SVM, Random forest regression, and Decision tree among these, SVM achieved the best accuracy of 88%.

The data had been made available by the king furs & fine jewelry website of Memphis [5] and the United States Geological Survey (USGS) African Diamond database [6] consists of various observations with features like each of the 4 C's of diamonds (Cut, Color, Clarity, and Carat) which affects the price. The data corpus is about the African diamond sale prices in Memphis during the last year, taken to estimate the market price of diamonds. The dataset has size 31700 and dimensionality 12 ('carat weight', 'cut', 'color', 'polish', 'Symmetry', clarity',' depth', 'table', 'length', 'width', 'height', 'price'). The typical data points look like these below.

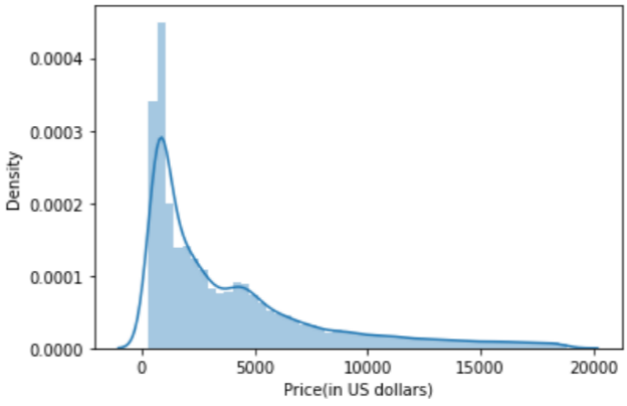
	Carat Weight	Cut	Color	Clarity	depth	table	Symmetry	Report	Polish	Length	Width	Height	Price
0	2.11	Ideal	H	SI1	61.5	55	VG	GIA	VG	3.95	3.98	2.43	18609
1	1.05	Very Good	E	VS1	59.8	61	G	GIA	VG	3.89	3.84	2.31	7666
2	0.91	Ideal	D	VS2	56.9	65	VG	GIA	VG	4.05	4.07	2.31	6224
3	1.01	Good	E	SI1	62.4	58	G	GIA	G	4.20	4.23	2.63	5161
4	0.92	Good	I	VS2	63.3	58	VG	GIA	VG	4.34	4.35	2.75	3679
5	2.51	Very Good	G	VS2	62.8	57	VG	GIA	VG	3.94	3.96	2.48	34361

CASE STUDY

A prediction problem was done by author Sanjoy Mondal [1], forecasting the market price of diamonds in the United States. As a part of the study Sanjoy has made use of various prediction algorithms such as linear regression, Decision tree, and Random Forest regressor. In that Random forest regressor got best performance, when comparison models with better metrics score with a relative error (RE) of at most 20% on average than other models. This was key to my project since I tried the same kinds of solutions.



Sanjoy did the histogram distribution plot to show the frequency price of how often each different value in a set of data occurs on his dataset.



TAKE-HOME DELIVERABLE

To obtain the deliverables, a lot of efforts are made and explained below. The data collected is fit into different algorithms after preprocessing and processing using python and the steps are described below.

For the assessment of this model, I used the metric relative error to assess the performance of the models. Firstly, I had split the entire dataset into 2 parts, 80% for training data and another 20% testing. Then model fitting by using predicting algorithms such as Linear regression, SVM, Decision tree and Random forest regressor.

```
In [8]: from sklearn.model_selection import train_test_split
        from sklearn import metrics
        from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 99)
```

Linear regression: Multiple Linear Regression there are more than one independent variables for the model to find the relationship of it with the dependent variable and best fit linear line and the optimal values of intercept and coefficients such that the error is minimized. Parameter used in linear regression model was “normalize”.

```
In [10]: from sklearn.linear_model import LinearRegression

In [11]: model_lr = LinearRegression()
         model_lr
Out[11]: LinearRegression()

In [12]: model_lr = model_lr.fit(X_train, y_train)
```

```
In [15]: pd.DataFrame(model_lr.coef_, X.columns, columns = ['Coeff'])
```

Out[15]:

	Coeff
Carat Weight	0.409965
Cut	0.005656
Color	-0.086976
Clarity	0.035339
Polish	-0.001597
Symmetry	-0.009218
Report	-0.008039
depth	-0.029786
table	-0.005071
Length	-0.058075
Width	0.015745
Height	0.046708

SVM: Basically, SVM finds a hyper-plane that creates a boundary between the types of each data item in the dataset in an N-dimensional space, where N is the number of features/attributes in the data to find the optimal hyper-plane to separate the data. Parameter used in SVM model was “kernel = rbf”.

```
In [19]: from sklearn.svm import SVR
         regressor = SVR(kernel="rbf")

In [20]: regressor = regressor.fit(X_train,y_train)
         y_pred = regressor.predict(X_test)
```

Decision tree regressor:
Decision tree regression observes features, breaks down the data set into smaller subsets and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Parameter used in Decision tree regressor model was “max_depth”.

```
In [29]: from sklearn.tree import DecisionTreeRegressor
         DecisionTreeRegModel = DecisionTreeRegressor()
         DecisionTreeRegModel = DecisionTreeRegModel.fit(X_train,y_train)
         y_pred = DecisionTreeRegModel.predict(X_test)
```

Random forest regressor:
Random Forest is based on the ensemble learning method and many Decision Trees. It finds the best decision tree among them with good parameters. Parameter used in Decision tree regressor model was “n_estimators”.

```
In [23]: from sklearn.ensemble import RandomForestRegressor
         reg_RF = RandomForestRegressor(n_estimators = 100, random_state = 0)

In [24]: reg_RF = reg_RF.fit(X_train,y_train)
         reg_RF

Out[24]: RandomForestRegressor(random_state=0)
```

Results of evaluation:
The R-squared error score is the metric used to evaluate the quality of the prediction models above after fitting the training data and testing it on the test data.

Prediction Models	R-squared Error score
Linear regression	0.79%
SVM regressor	0.66%
Decision tree regression	0.83%
Random forest regressor	0.95%

When compared to other models, the random forest regressor model had the lowest relative error of 0.05%.

```
In [30]: y_pred = reg_RF.predict(X_test)

In [31]: print("R^2:",metrics.r2_score(y_pred, y_test))
         R^2: 0.9548486766477772

In [32]: print("MAE:",metrics.mean_absolute_error(y_pred, y_test))
         print("MSE:",metrics.mean_squared_error(y_pred, y_test))
         print("RMSE:",np.sqrt(metrics.mean_squared_error(y_pred, y_test)))
         MAE: 0.010418321075482634
         MSE: 0.0004718056149741132
         RMSE: 0.021721086873683675
```

REFERENCES

[1] Grover J & Louden C (2021, April 11). 100% ML: Diamond price prediction using machine learning, Python, SVM, KNN, Neural Networks. Five Step Guide. https://fivestepguide.com/technology/machine-learning/diamond-price-prediction-using-machine-learning/

[2] Alsuraihi W, Al-hazmi E, Bawazeer K, & Alghamdi H (2020, March 1). Machine learning algorithms for diamond price prediction: Proceedings of the 2020 2nd International Conference on Image, video and Signal Processing. ACM other conferences. https://dl.acm.org/doi/10.1145/3388818.3393715

[3] Abirami R., & Agniswar (2022, December 1). Automated diamond price prediction using machine learning. SRM University AP. https://srmap.edu.in/wp-content/uploads/2021/12/Automated-Diamond-Price-Prediction-Using-Machine-Learning.pdf?x81859

[4] Mihir H, Patel M, Jani S, & Gajjar R. (2021). Diamond price prediction using machine learning. 2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4). https://doi.org/10.1109/c2i454156.2021.9689412

[5] https://www.kingfursandfinejewelry.com/design-your-own-ring/choose-your-diamond

[6] https://pubs.er.usgs.gov/publication/ofr20181088

[7] https://www.kaggle.com/code/sanjoymondal0/diamond-price-prediction-regression#Linear-Regression

[8] Code access link FDS Final Term Project code.html