CENG3522: Applied Machine Learning (Spring 2023) Midterm

Mehmet Kadri Gofralılar - 180709005

Q1. (5 points) State the name of your project and full names (first name/last name) of your team Members.

Project Name: Gemstone Price Prediction

Mehmet Kadri Gofralılar

Mehmet Enes Kızılkaya

İrem Kızılırmak

Q2. (20 points) Group the features in your dataset into Categorical/ordinal, Categorical/nominal, Numeric/ratio-scaled, Numeric/interval-scales.

Categorical (Ordinal) : Cut, Color, Clarity

Numeric (Ratio-scaled): Unnamed, Carat, Depth, Table, X, Y, Z, Price

Q3. (15 points) Describe what is the learning scheme (e.g. regression) for your project problem. Justify your answer by describing a business or personal use (i.e. how a company or person can use your model?).

Regression is the main learning scheme, since the target variable is numeric. If the results of the regression are insufficient, discretization and classification could be tried as well. After inspecting the dataset furthermore, we realized that there are outliers which might decrease the accuracy since predicting a specific value is much harder than predicting a discretized class value for outliers. So, we decided to apply classification methods for our dataset if regression fails.

From a seller's point of view, estimating the price of a “cubic zirconia” is important in order to sell the product with the highest reasonable price. From the customer’s point of view, estimating the price of a “cubic zirconia” is important in order to not get overcharged.

If a company/person needs to know the price of a “cubic zirconia” whose features are known, running the model with required inputs gives them the estimated price. But since regression might have a bigger chance of failure because of the lack of instances with similar features, predicting the interval of the price as class might be better to increase the accuracy.

Q4. (30 points) Provide your decisions on handling at least five(5) features in your dataset using following table:

	Feature	**Would you keep? (Yes/No)**	**(If kept) What transformation(s)/ derivation(s) would you do?**	**Why do you keep(or not keep)? (If kept) Why do this transformation/derivation?**
1	Unnamed	No		Redundant index variable
2	Cut	Yes	Label Encoding	Machine learning algorithms need numeric variables, therefore categorical variables should be transformed. Since this variable is an ordinal variable, using label encoding is better compared to one hot encoding.
3	Colour	Yes	Label Encoding	Machine learning algorithms need numeric variables, therefore categorical variables should be transformed. Since this variable is an ordinal variable, using label encoding is better compared to one hot encoding.
4	Clarity	Yes	Label Encoding	Machine learning algorithms need numeric variables, therefore categorical variables should be transformed. Since this variable is an ordinal variable, using label encoding is better compared to one hot encoding.
5	X, Y, Z -> Volume	Yes	Volume ≈ z/mean(x,y)	A new feature named “Volume” can be derived using X, Y and Z features with the hope of creating a more meaningful single feature for predicting the price value for the gemstone.
6	Carat, Volume	Yes	Density = Carat/Volume	A new feature named “Density” can be derived using Carat and Volume features with the hope of creating a more meaningful single feature for predicting the price for the gemstone if the volume feature is insufficient on its own.

Q5. (15 points) Name at least three(3) algorithms you can use for your project based on the scheme stated in Q3.

Regression algorithms: Lasso/Ridge Regression (could help for normalization), k-Nearest Neighbors, Decision Tree, Random Forest, Support Vector Machine

If regression is insufficient, **classification algorithms:** Support Vector Machine, K-Nearest Neighbors, Logistic Regression, Naive Bayes Classifier, Decision Tree, Random Forest

Q6. (15 points) Describe the metrics you will use in assessing the performance of your learned models. Discuss strength and shortcomings of individual metrics.

(Regression) 1. Mean Squared Error (MSE): Calculates the average squared distance between predicted and original values of the target variable.

Strengths:

- It highlights outliers that have a greater effect on model's performance.
- It is used more in optimization algorithms for model training.

Shortcomings:

- It is hard to interpret since it squares the errors.
- Not robust, affected by outliers easily.

(Regression) 2. R-squared (R^2): Shows how close the data are to the fitted regression line with value between 0-1.

Strengths:

- It shows how well the model fits the data.
- It is easy to interpret. Higher value means better fit.

Shortcomings:

- Does not show the magnitude or direction of prediction errors.
- Comparing R-squared values between models with different numbers of predictors can be misleading.

(Classification) 1. Accuracy: Measures the proportion of correctly predicted instances out of the total number of instances. It provides a general overview of the model's performance.

Strengths:

- It is easy to understand and interpret.
- It works well when the classes are balanced.

Shortcomings:

- Accuracy can be misleading when classes are imbalanced. Predicting the majority class for every value could result in high “accuracy”, when there is no learning/prediction, just underfitting.

(Classification) 2. F1 Score: Is the harmonic mean of precision and recall. It combines both metrics to provide a balanced measure of the model's performance.

Strengths:

- It is useful when there is an imbalanced class distribution or when false predictions are important.

Shortcomings:

- Does not consider true negatives.