Introduction to Data Science

Fall 2023

Total marks 110

Due date: 28th November, 2023, 11:59 PM

Assignment 4

Instructions:

• All questions must be answered within a single notebook or .py file.
• Follow the file naming conventions: Name your submission file as RollNo.ipynb or RollNo.py (e.g., i22_xxxx.ipynb, where xxxx is your Roll Number).
• Use headings to distinguish each question in the notebook.
• Late submissions will not be accepted and will be given a zero.
• AI-generated content is prohibited. Detection of such content will lead to a zero score.

1. **Regression Analysis [60 Marks]**

   Problem Statement:

   As a data analyst hired by a real estate company, your objective is to utilise statistical insights to support decision-making in property transactions, acquisitions, leasing, and management. The dataset provided is extracted from **Zameen.com**, a prominent property portal in Pakistan.

   The real estate company's clients, located nationwide, have properties they wish to sell but require a specific pricing strategy. Your responsibility is to streamline this process through **regression analysis**. The goal is to develop a machine learning model capable of accurately predicting property prices based on various attributes, including area and the number of rooms.

   This task will be executed in four phases:

   Choose a **specific city *or* province** from the dataset as the focal point for regression analysis and filter the data accordingly.

   Data Pre-Processing (10 Marks):

   - Ensure the validation and correction of any inconsistent data formats.
   - Identify and address missing values in the dataset.
   - Identify and manage potential outliers within the data.

   Exploratory Data Analysis (EDA) (10 Marks):

   - What is the overall correlation structure within the dataset? Are there any notable high or low correlations between variables?

- Is there a correlation between the number of properties listed by an agent or agency and the average property price?

Feature Engineering (15 Marks):

- Compute a new column indicating the price per square meter, considering that the **area** column is in square meters.
- Derive additional temporal features, such as month, quarter, or day of the week, from the **date_added** column.
- Standardise the numerical variables using a suitable standardisation technique.
- Encode the categorical variables using an appropriate encoding method.

Note: To enhance model accuracy, eliminate any unnecessary columns.

Model Training (15 Marks):

- Divide the data into training and testing sets, selecting a suitable ratio for the split.
- Select an appropriate regression model and train it on the divided data. Ensure to **perform hyperparameter tuning**, whether using the kitchen sink method, an exhaustive search approach, or any suitable technique.

Model Evaluation (10 Marks):

It is crucial to validate the accuracy of your trained machine learning model. This can be accomplished by producing the following:

- **Mean Absolute Error:** It is the average of absolute differences between the predicted and actual values – ranges from zero to infinity, with lower values indicating better accuracy.
- **Mean Squared Error:** It is the average of squared differences between the predicted and actual values – ranges from zero to infinity, with lower values indicating better accuracy.
- **Root Mean Squared Error:** It is the square root of the mean squared error, which brings the error metric back to the original scale of the target variable – ranges from zero to infinity, with lower values indicating better accuracy.
- **Mean Absolute Percentage Error:** It is the average of the absolute percentage differences between the predicted and actual values – ranges from 0% to 100%, with lower values indicating better accuracy.

Bonus (10 Marks):

In regression analysis, the goal is to build a model that predicts a target variable based on one or more predictor variables. When the number of predictor variables (features or dimensions) is large, it may lead to the **curse of dimensionality**.

Your responsibility is to assess whether your regression model is affected by the curse of dimensionality. To achieve this, you can systematically investigate methods such as cross-validation and regularisation. If indeed there is an issue with the curse of dimensionality, your responsibility is to tackle it using a suitable approach, such as feature selection, dimensionality reduction, or regularisation.

Read More: How to break the "Curse of Dimensionality"?

2. **Predicting Tax Fraud Risk Using Decision Trees [50 Marks]**

Objective:

In the realm of combating potential tax fraud, the primary aim of this question is to develop a predictive model employing decision trees. This model will be constructed based on a dataset encompassing various attributes related to individuals. The dataset includes the following features: (Company_Data.csv)

- Undergrad: A binary indicator denoting whether an individual is under-graduated or not.
- Marital Status: The marital status of the individual.
- Taxable Income: The amount of taxable income, serves as an indicator of the individual's tax liability to the government.
- Work Experience: The number of years of work experience of the individual.
- Urban: A binary indicator specifying whether the person resides in an urban area.
- Taxable Income

Target Variable:

The target variable for classification is defined as follows:

- Individuals with a taxable income less than or equal to $30,000 are labeled as "Risky."
- Individuals with a taxable income greater than $30,000 are labeled as "Good."

Model Development:

The goal is to construct a decision tree model that effectively classifies individuals into the aforementioned categories based on their attributes. The model should be particularly adept at identifying the risk of tax fraud, with a focus on individuals possessing lower taxable incomes.

Model Evaluation:

The success of the model will be assessed through key metrics, including accuracy, precision, recall, and F1 score. These metrics will provide a comprehensive understanding of the model's performance in distinguishing between "Risky" and "Good" individuals.

Feature Engineering Steps:

To prepare the data for modeling, the following feature engineering steps will be implemented:

- **Categorical Variable Handling**: Utilize the pandas get_dummies function to convert categorical variables (Undergrad, Marital Status, and Urban) into dummy or indicator variables.

  Reference: pandas.get_dummies

- **Target Variable Transformation**: Define the target variable assuming taxable income <= $30,000 as "Risky=0" and others as "Good=1."

- **Feature Scaling**: Normalize the data by scaling the "Work Experience" and "City Population" features to a specified range using the MinMaxScaler from scikit-learn.

  Reference: sklearn.preprocessing.MinMaxScaler

Data Splitting:

Split the dataset into training and testing sets using the train_test_split function from scikit-learn.

Reference: sklearn.model_selection.train_test_split

Model Training and Evaluation:

After training the model on the training data, evaluate its performance on the test set. Display the classification report to provide a detailed overview of the model's precision, recall, and F1 score.

Reference: sklearn.metrics.classification_report

Model Improvement:

Explore strategies to enhance the accuracy of the model. Experiment with hyperparameter tuning, feature selection, or other techniques to optimize the model's predictive capabilities.

Happy Coding 🙂