CECS 456 Machine Learning

Spring 2023 – Final Project

Deliverables

1. **Source code** on GitHub repo
2. **Presentation slides** (upload on Canvas)
3. **Project report** in Microsoft word file (upload on Canvas)

Submission instruction and deadline: Please zip “presentation slide” and “project report” and upload them on Canvas by 04/26/2023.

Note: Only one submission is required from each group and Group leader should upload the files on Canvas.

Group 1: Productivity Prediction of Garment Employees

1. Perform exploratory data analysis and feature selection
    a. Plot the variable distribution using box plot, histogram, and provide a summary of your understanding.
    b. Provide an insight on the outcome of correlation matrix
2. Perform feature extraction using Principal Component Analysis (PCA).
3. Develop a machine learning model to predict the “productivity performance” of the employees and compare the “performance metrics” of the algorithms. Summarize your key findings.
    a. Please consider the problem statement as **Regression**
    b. First begin with Linear regression, then evaluate the performance metrics of logistic regression, SVM regressor, and Random Forest regressor.
    c. Plot the relevant graphs
4. Recommend two approaches for increasing the productivity of the employees.

Dataset

https://archive.ics.uci.edu/ml/datasets/Productivity+Prediction+of+Garment+Employees

Garrett Chavez
Nathan Pry
Jose Fuentes
Richard Huang
William Pham
Om Shah
Deep Meghani
Om Kakadiya

Group 2: Early-stage diabetes risk prediction

1. Perform exploratory data analysis (EDA), for example histogram of the features, boxplot, and part from this you are encouraged to explore EDA and plot relevant graphs.
2. Develop a machine learning model to identify if the patient is newly diabetic or would be diabetic patient using the dataset.
    a. Compare and tabulate the performance metrics (confusion matrix) of logistic regression, random forest classifier, and support vector machines classifier on the problem statement.
    b. Plot ROC and AUC
3. What do you think is the most prominent features responsible for diabetic patients? You may perform feature ranking using suitable algorithm.

Dataset

https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.

Kenneth Valero
Carlsean Claricia
Brian Vu

Brian Tran
Sean iida
Abhay Solanki
Sierra Harris

Group 3: Online Shoppers Intention

The goal is to explain what actions of customer browsing on an e-commerce website will contribute to that customer buying a product. You will be using clustering and classification algorithms to make predictive models around shoppers' intentions. OBJ

1. Perform exploratory data analysis and data pre-processing.
    a. There are missing values in the dataset. One way of handlining missing values is to perform imputation, but in certain cases we can also drop them. How you want to treat the missing values? Handle categorical variables using one-hot encoding.
    b. Plot the histogram of the features and boxplot for outlier detection,
    c. Plot the correlation matrix for the features.
    d. Provide a list of top 10 most important features.
2. The dataset contains 18 features. Your task is to use Principal Component Analysis (reduce the number of variables of a data set).
3. Design a machine learning algorithm to identify the behavior of customers if they are going to purchase the product or not.
    a. Use k-nearest neighbor, Naïve Bayes, logistic regression, SVM, and Random Forest classification algorithm to create online shopper intention (target variable: Is_Revenue).
    b. Compare the performance metrics of the classification algorithms.
    c. Plot ROC graph for the algorithms
4. Suggested a few possible ways to attract more customers to finish with purchasing.

Dataset

https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

Phiona Nicole Tumbaga
Alberto Perez
Joey Rice <3
Eric Chhour
Jonathan Santos
Hadi Al Lawati
Abdullah Al Nahwi

Group 4: Automobile Price Prediction

Task: Regression

Tasks: Predicting the price of a car based on its technical specifications such as engine size, horsepower, curb weight, etc.

Hints and Insights:

1. When using linear regression, which kind of regression technique works best. Simple, multiple or polynomial regression?

2. Take into consideration B**ias vs Variance** tradeoff.

3. Apply regularization in order to prevent Overfitting (consider L1, L2 Regularization techniques and provide reasoning on what basis you have chosen this technique).

Instructions/Directions:

1. Perform exploratory data analysis and feature selection
2. Visualization

    2.1. Plot the variable distribution using box plot, histogram, and provide a summary of your understanding

    2.2. Provide an insight on the outcome of the correlation matrix.

    2.3. Plot the relevant graphs

3. Perform feature extraction using Principal Component Analysis (PCA).
4. For Model selection, perform K-Fold Cross Validation on the regression models used.
5. Design a regression model (use at least 5 regression models) and perform hyperparameter tuning and document the results.

Dataset: https://archive.ics.uci.edu/ml/datasets/Automobile

Mayank Tamakuwala
Dhruv Gorasiya
Keshav Mehta
Ayush Patel
Marshall Keopong
Sean Collins
Christopher Vasquez

Group 5: South German Credit

1. Perform exploratory data analysis (EDA) using data visualization, for example, histogram of the features, boxplot, and apart from this you are encouraged to explore EDA and plot relevant graphs.
    a. Identify the outliers in the dataset
    b. Plot the correlation matrix for the dataset.
    c. Plot the graphical distribution for the variables
2. Identify the optimal number of clusters in the dataset. You may want to compare silhouette and elbow method.
3. Use k-means algorithm for creating the clusters.
    a. Credit amount vs age
    b. Credit amount vs duration
    c. Age vs duration
4. Interpret each of the clusters in question 3.

5. Use HDBSCAN to perform hierarchical clustering and plot the dendrogram.
6. Explore various performance metrics of clustering algorithm and compare them.

Dataset

https://archive.ics.uci.edu/ml/datasets/South+German+Credit

David Nguyen
Simon Ngo
Alan Marin
Austin Elizondo
Janelle Chan
Jason Barber
Matthew Chung

Group 6: Productivity Prediction of Garment Employees

1. Perform exploratory data analysis and feature selection
2. Plot the variable distribution using box plot, histogram, and provide a summary of your understanding.
3. Provide an insight on the outcome of correlation matrix
4. Perform feature extraction using Principal Component Analysis (PCA).
5. Develop a machine learning model to predict the “productivity performance” of the employees and compare the “performance metrics” of the algorithms. Summarize your key findings.
6. Please consider the problem statement as **Classification**
7. Evaluate the performance metrics of logistic regression, SVM classifier, Naïve Bayes classifier, and Random Forest classifier.
8. Plot the relevant graphs, for example ROC, AUC, etc.
9. Recommend at least two approaches for increasing the productivity of the employees.

Dataset

https://archive.ics.uci.edu/ml/datasets/Productivity+Prediction+of+Garment+Employees

James Cho
Ryan Randazzo
Kevin Cordray
Bryant Lam
James Ta
J Garcia
Hai Jiao Cui

Group 7: Real estate valuation

1. Perform exploratory data analysis (EDA), for example histogram of the features, boxplot, and apart from this you are encouraged to explore EDA and plot relevant graphs.
    a. Identify the outliers in the dataset
    b. Plot the correlation matrix for the dataset.
    c. Plot the graphical distribution for the variables
2. Perform feature extraction using Principal Component Analysis (PCA) and do the necessary data pre-processing, for example, feature scaling.
3. Design a machine learning model to evaluate the real estate price.
    a. Begin with linear regression, logistic regression, and then support vector machine regressor.
    b. Use random forest algorithm and perform hyper-parameter tuning to see the model improvement. Document the results.
    c. Compare and tabulate the performance metrics (Mean Absolute Error, R-Squared) of the regression algorithms used.

Dataset

https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set

Faizan Zafar
Jason Morales
Chris Dao
Bao Nguyen
Vu Phan
Tina Vu
Gavin Lampton

Group 8: Carbon-Monoxide Prediction for Air-Quality Dataset

Task: Regression

Tasks: Predict CO concentration: One of the main pollutants measured in this dataset is Carbon Monoxide (CO). A regression model is to be trained to predict the CO concentration based on the other measured variables such as temperature, humidity, and other pollutants.

Data Set Information:

The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was located on the field in a significantly polluted area, at road level,within an Italian city.

Data were recorded from March 2004 to February 2005 (one year)representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer.

Evidences of cross-sensitivities as well as both concept and sensor drifts are present eventually affecting sensors concentration estimation capabilities.

Missing values are tagged with -200 value.

Instructions/Directions:

The Air Quality dataset contains hourly measurements of air quality from an Italian city.

1. Perform exploratory data analysis and feature selection
2. Visualization

    2.1. Plot the variable distribution using box plot, histogram, and provide a summary of your understanding

    2.2. Provide an insight on the outcome of the correlation matrix.

    2.3. Plot the relevant graphs

3. Perform feature extraction using Principal Component Analysis (PCA).
4. For Model selection, perform K-Fold Cross Validation on the regression models used.
5. Design a regression model (use at least 5 regression models) and perform hyperparameter tuning and document the results.

Dataset: https://archive.ics.uci.edu/ml/datasets/Air+Quality

Nathan Lai
Sarthak Nagrani
Brenden Inhelder
Onyedikachi Benjamin Okenwa
Williams Nguyen
Sopheak Chim
Dylan Ramos

Group 9: Breast Cancer Prediction

Task: Classification

Dataset Description

This data set includes 201 instances of one class and 85 instances of another class. The instances are described by 9 attributes, some of which are linear, and some are nominal.

Tasks:

Predict if the patient has breast cancer or not (binary classification problem)

Instructions/Directions:

1. Perform exploratory data analysis and feature selection
2. Visualization

    2.1. Plot the variable distribution using box plot, histogram, and provide a summary of your understanding.

    2.2. Provide an insight on the outcome of correlation matrix

3. Perform feature extraction using Principal Component Analysis (PCA).

4. Plot the confusion matrix.
5. Evaluate the performance metrics of logistic regression, SVM classifier, Naïve Bayes classifier, and Random Forest classifier.
6. Plot the relevant graphs, for example ROC, AUC, etc.
7. Use Artificial Neural Network (Deep Learning Method) and compare the accuracy with traditional Machine Learning Models. Write down your observations.

Dataset: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer

Phong Cao
Peter Pham
Jeremy Lin
Omar lee
Maximillian Gutierrez
Anthony Sanchez
John Messina

Group 10: Analyzing Tourist Satisfactory based on feedback provided on East Asia

Task: Clustering

Tasks: Cluster the reviews based on user feedback and identify the attractions of East Asia based on Attributes presented

Description: This data set is populated by crawling TripAdvisor.com. Reviews on destinations in 10 categories mentioned across East Asia are considered. Each traveler rating is mapped as Excellent (4), Very Good (3), Average (2), Poor (1), and Terrible (0) and average rating is used against each category per user.

Instructions/Directions:

1. Perform exploratory data analysis (EDA) using data visualization, for example, histogram of the features, boxplot, and apart from this you are encouraged to explore EDA and plot relevant graphs.

   1.1 Identify the outliers in the dataset

   1.2 Plot the correlation matrix for the dataset.

   1.3 Plot the graphical distribution for the variables

2. Identify the optimal number of clusters in the dataset.

   2.1. You may want to compare silhouette and elbow method.

3. Use k-means algorithm for creating the clusters.

   3.1 Interpret each of the clusters in question 3.

4. Use HDBSCAN to perform hierarchical clustering and plot the dendrogram.
5. Compare the results of various clustering algorithms

Dataset: https://archive.ics.uci.edu/ml/datasets/Travel+Reviews

Haley Nguyen
Joaquin Alonzo
Joshua Quibin
Allison Barajas
Gustavo Pech
Manav Dillon
Jedidiah Shank

Group 11: Skin Cancer Prediction on Dermatology Dataset

Task: Classification

Tasks:

Classifying skin diseases into two categories, "benign" and "malignant", based on the diagnosis provided in the dataset. This can be a binary classification task where the goal is to predict whether the skin disease is benign or malignant.

Data Set Information:

This database contains 34 attributes, 33 of which are linear valued and one of them is nominal.

The differential diagnosis of erythemato-squamous diseases is a real problem in dermatology. They all share the clinical features of erythema and scaling, with very little differences.

The diseases in this group are

1. psoriasis,
2. seboreic dermatitis,
3. lichen planus,
4. pityriasis rosea,
5. cronic dermatitis, and
6. pityriasis rubra pilaris.

Instructions/Directions:

1. Perform exploratory data analysis and feature selection
2. Visualization

2.1. Plot the variable distribution using box plot, histogram, and provide a summary of your understanding.

2.2. Provide an insight on the outcome of correlation matrix

3. Perform feature extraction using Principal Component Analysis (PCA).
4. Plot the confusion matrix.
5. Evaluate the performance metrics of logistic regression, SVM classifier, Naïve Bayes classifier, and Random Forest classifier.
6. Plot the relevant graphs, for example ROC, AUC, etc.

7. Use Artificial Neural Network (Deep Learning Method) and compare the accuracy with traditional Machine Learning Models. Write down your observations

Dataset: https://archive.ics.uci.edu/ml/datasets/Dermatology

Jett Sonoda
Darius Koroni
Nhi Pham
David De Girolamo
Christopher Pineda
Brianna Soriano
Corbin Marino

Group 12: Nitrogen Oxide Concentration Prediction in air quality dataset

Task: Regression

Tasks:

Predicting NOx concentration: A pollutant measured in this dataset is Nitrogen Oxides (NOx). A regression model is to be trained to predict the NOx concentration based on the other variable/features.

Data Set Information:

The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was located on the field in a significantly polluted area, at road level,within an Italian city.

Data were recorded from March 2004 to February 2005 (one year)representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer.

Evidences of cross-sensitivities as well as both concept and sensor drifts are present eventually affecting sensors concentration estimation capabilities.

Missing values are tagged with -200 value.

Instructions/Directions:

The Air Quality dataset contains hourly measurements of air quality from an Italian city.

1. Perform exploratory data analysis and feature selection
2. Visualization

2.1. Plot the variable distribution using box plot, histogram, and provide a summary of your understanding

2.2. Provide an insight on the outcome of the correlation matrix.

2.3. Plot the relevant graphs

3. Perform feature extraction using Principal Component Analysis (PCA).

4. For Model selection, perform K-Fold Cross Validation on the regression models used.
5. Design a regression model (use at least 5 regression models) and perform hyperparameter tuning, and document the results.

Dataset: https://archive.ics.uci.edu/ml/datasets/Air+Quality

Ian Escalante
Andrew Bae
Francisco Rivera
Anh Huynh
Carlos Bordallo
Joshua Hicks
Aster Lee

Group 13: Student behavioral patterns in Online Learning platform

Task: Clustering

Tasks:

Clustering student behavior based on their test scores

Instructions/Directions:

1. Perform exploratory data analysis (EDA) using data visualization, for example, histogram of the features, boxplot, and apart from this you are encouraged to explore EDA and plot relevant graphs.

    1.1 Identify the outliers in the dataset

    1.2 Plot the correlation matrix for the dataset.

    1.3 Plot the graphical distribution for the variables

2. Identify the optimal number of clusters in the dataset.

    2.1. You may want to compare silhouette and elbow method.

3. Use k-means algorithm for creating the clusters.

    3.1 Interpret each of the clusters in question 3.

4. Use HDBSCAN to perform hierarchical clustering and plot the dendrogram.
5. Compare the results of various clustering algorithms.

Dataset:https://archive.ics.uci.edu/ml/datasets/Educational+Process+Mining+%28EPM%29%3A+A+Learning+Analytics+Data+Set

Jovanni Garcia
Joshua Gherman

Jason Jitsiripol
chase aufmann
Matthew Kriesel
Ryan Gieg
Ryo Fujimura

Group 14: Benzene Concentration Prediction in air-quality dataset

Task: Regression

Tasks:

Predicting Benzene concentration: Benzene is a carcinogenic pollutant that is measured in this dataset. A regression model is to be trained to predict the Benzene concentration based on the other variables.

Data Set Information:

The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was located on the field in a significantly polluted area, at road level,within an Italian city.

Data were recorded from March 2004 to February 2005 (one year)representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer.

Evidences of cross-sensitivities as well as both concept and sensor drifts are present eventually affecting sensors concentration estimation capabilities.

Missing values are tagged with -200 value.

Instructions/Directions:

The Air Quality dataset contains hourly measurements of air quality from an Italian city.

1. Perform exploratory data analysis and feature selection
2. Visualization

    2.1. Plot the variable distribution using box plot, histogram, and provide a summary of your understanding

    2.2. Provide an insight on the outcome of the correlation matrix.

    2.3. Plot the relevant graphs

3. Perform feature extraction using Principal Component Analysis (PCA).
4. For Model selection, perform K-Fold Cross Validation on the regression models used.
5. Design a regression model (use at least 5 regression models) and perform hyperparameter tuning, and document the results.

Dataset: [UCI Machine Learning Repository: Air Quality Data Set]

Ryley Benavides

David Shamis
Noah Daniels
Pranik Pant
Dhruv Savla