/Users/apple/Library/Containers/com.microsoft.Outlook/Data/Library/Caches/Signatures/signature_2105988401

**Principles of Data Science and Computing Systems**

**10204280**

**Assignment Title**

Application of Data Science Life Cycle

Assignment 2

**Submitted to**

Dr. Heba Alawneh

**Submitted by**

Yousef AbuAli

2022/2023(Fall Semester)

1. Introduction:
   1. Type of Machine Learning:

|  |  |  |
| --- | --- | --- |
|  | **Supervised Learning** | **Unsupervised Learning** |
| **Definition** | A type of machine learning, it basically training a model on a labelled dataset, which means to provide it with input data and correct outputs. Which later on can be used by the machine to predict new, unseen data. | A type of machine learning, it works by training a data on unlabelled dataset, which means that the model is only provided with input data without correct output. Its goal is to make the model discover common patterns in the data by itself. |
| **Applications (use)** | * Image classification * Finance * Marketing * Robotics | * Clustering * Outlier detection: identifying unusual data points. * Data visualization: to represent data in a 2D or 3D to make it easier to understand. |
| **Strengths** | * High accuracy, specially when trained on big, good quality data. * High-performance * Flexibility, as it can be applied to many tasks | * It doesn’t need labelled data. * Can discover hidden patterns. * Able to identify outliers. |
| **Limitations** | * It requires labelled data. * Limited to the training data, which means it makes predictions only related to it. * Overfitting due to memorizing the training data and perform poorly on unseen data. | * Hard to evaluate because there is no specific outcome to compare. * Requires big amount of data to learn. * Does not provide feedback such as the accuracy, which makes it difficult to make improvements. |
| **Common Algorithms** | * Linear regression * Logistic regression * Decision trees * k-Nearest Neighbors(k-NN) | * Clustering: such as K-means clustering, Hierarchical clustering. * Association |

* 1. Supervised machine learning:

Classification and regression are both types of a supervised machine learning algorithms.

Regression is used with continuous values, for example, while predicting a price of a house there will be no specific outcomes. The main goal for regression is to make the model predict new data accurately.

On the other hand, Classification is used with discrete data, going back to the example of houses you can say that a price of a house is either cheap, moderate, or expensive. Therefore, we only have a specific outcome. The main goal for classification is to make the model predict the class of the new data points.

Moreover, other main differences of classification and regression is that in classification the input values could be either continuous or discrete. Whereas the input variable in regression are only continuous values.

In conclusion, the type of inputs and outputs are main differences of classification and regression.

|  |  |  |
| --- | --- | --- |
|  | **Classification** | **Regression** |
| **Type of Learning** | supervised | supervised |
| **Applications (use)** | * Fraud detection * Medical diagnosis * Text classification: is it an email a spam or not? | * Financial forecasting: such as predicting the stock prices. * Weather forecasting: such as predicting the temperature. * Crime forecasting: predicting the likelihood of a crime depending on factors, ex, poverty. |
| **Strengths** | * Managing categorical data * Managing large datasets * Used for feature selection. * Able to identify labels or classes for new data | * Able to handle largen number of data as inputs * Able to work with small samples, such as the linear regression, and it will be able to produce good results. * Able to handle continuous data |
| **Limitations** | * Sensitive to noise and outliers * Assumes independence between data and this is not the case with real-world problems. * It needs labelled data for training. | * Sensitive to outliers because it can affect the model estimates. * Requires numerical variables to make predictions. * a linear relationship is assumed between dependent and independent variables, and in reality it might not be a linear relationship. |
| **Common Algorithms** | * Decision Trees * k-Nearest Neighbors (k-NN) * Naive Bayes | * Linear regression * Neural networks |

* 1. Computing systems

Both Microsoft azure and Amazon web server (AWS) are considered as cloud computing services. Both of these uses different hardware systems to provide the cloud computing services. In other words, Microsoft azure and AWS give their users the opportunity to rent services and computing resources instead of buying and maintaining their own hardware devices. In addition, there are some mutual services provided by both these platforms, such as storage, databases, virtual machines, networking, and security.

Differences between Microsoft azure and Amazon Web Server (AWS) based on the structure, GPU, CPU, type of the storage devices being used.

1. Structure:

* AWS uses a combination of resources that are both physical and virtualized, and they both provide the cloud computing service. For example, a combination of servers(physical), virtual machines, as well as containers.
* Azure uses an approach that is similar to the one used in AWS, it uses a combination of virtualized and physical resources, it includes serverless computing, virtual machines, as well as containers.

1. GPU

Definition:

a GPU is a processer that made to handle complex math operations that are needed for difficult tasks, such as crating images, videos, and animations.

GPU instances are considered as a virtual machine in a cloud computing environment which are designed to perform calculations in large numbers very fast.

* Both AWS and azure offers a GPU instances that are used for scientific computing, machine learning, as well as other computing workloads that needs high performance.

1. CPU

Definition:

A CPU is a processer that is used for general purposes, and it is used to perform multiple tasks, such as running operating systems, executing programs.

CPU instances are considered as virtual machines in cloud computing that were designed to perform wide range of tasks.

* Both AWS and azure provides CPU instances in wide range. It includes memory-optimized, general-purpose, and compute-optimized instances. The previous instances are based on processors, such as intel Xeon.

1. Storage devices:

* AWS provides multiple storage devices, such as network-attached storage, local storage, as well as storage area networks. In addition, it provides multiple storage services. For object storage it provides the azure blob storage, for block storage it provides the azure disk storage, and for file storage it provides the azure file storage.

1. Memory:

The two companies provides instances that contains various amount of memory, for example the general-purpose instances contains 2-16 GB of memory space, as well as the memory-optimized instances that contains up to 672 GB of memory space.

1. Security:

Both companies provide multiple services for security, such as the web application firewalls(WAFs), access and identity management(IAM), as well as the key management services(KMS).

1. Methodology:
   1. data science life cycle
2. Understanding the problem:

The problem is to analyze a certain data that is related to a food trading company. This data contains some information about the product that were sold last month in multiple stores in the country. And it is being used to help the risk management system of a financial company make decisions. The goal of the analyses is to clean the data and make sense of it, as well as performing the appropriate prediction, and then present the outcome(results) to customers who doesn’t have experience in this field.

1. Gathering the data(data mining):

I downloaded the data using CSV function from pandas library. I got information about each one of the features, such as knowing what each one represent, knowing the number of missing values in each feature, knowing the maximum and minimum values in each feature, finding wither there is correlation between the different features.

1. Data cleaning:

The data collected is cleaned by removing(dropping) irrelevant features as well as features that might have high correlation with another one. In addition, I filled the features that contains missing values with the appropriate functions. Moreover, I fixed the features that contains duplicate data to avoid outliers or errors in the data. I also, used two types of normalization techniques, such as min-max normalization, and z-score normalization.

1. Data exploration:

After the data was cleaned, I am supposed to gain some insight’s about the data. In this step I used histograms as visualization for each feature in the data.

1. Modeling data:

In this step I used supervised machine learning approaches, such as

k-Nearest Neighbors (k-NN) and linear regression in order to train the model with the cleaned dataset. In addition, I created some evaluation methods to evaluate the performance of the model.

|  |  |
| --- | --- |
| **The Data Life Cycle Stage** | **The importance of the given Stage toward your project lifecycle** |
| Business understanding | This stage helps in understanding the objectives as well as the general problems of the project. As well as it helps it setting a clear goal that were going to work to achieve. |
| Data mining | In this stage the data is collected, and prepared. If the data is accurate and relevant to each other it would be crucial for the project in order to succeed. |
| Data cleaning | In this step, the data is being cleaned and prepared to be ready for the next steps. This step is important in order to remove outliers, errors, as well as filling missing data. This step help in making sure that the data is reliable. |
| Data exploration | In this stage the data is being examined to understand the general characteristics of the data. It helps to understand wither there are issues in the data. |
| Feature engineering | This step is done by making a new feature or changing a certain feature in order to improve the data quality and make it better for analysis. It helps in improving the performance as well as the accuracy of the model. |
| Predictive modelling | In this stage machine learning is applied in order to make predictions from the data. This step mainly depends on the previous ones in order to produce valuable and accurate information’s. |
| Data visualization | In this step, the result of the analysis is used in order to represent them in an easy way for the non-technical stakeholders to understand. Therefore, this step helps to deliver the obtained results from the project to help the stakeholders make better decisions. |

Visualizing the features in the dataset using the Histogram:

|  |  |
| --- | --- |
| I\_W column. It shows that the most common wight was 12.5 |  |
| I\_Fat\_C histogram shows that the most frequent value is low fat product. | Chart, bar chart  Description automatically generated |
| I\_Vis histogram shows that the most common percentage for the total display area of all items in the store was between zero and 0.05. | Chart, histogram  Description automatically generated |
| I\_category histogram shows that the most common categories were fruits and vegetables as well as snack food. |  |
| I\_MRP\_US histogram shows that the most common maximum retail price for products in dollars was around 150. |  |
| O\_Establ\_Y histogram shows that most stores were established in 1995 and the year 2005. |  |
| I\_Recalled histogram shows that most of the items were not recalled. |  |
| O\_Loc\_T histogram shows that most of the stores were located in tier3. |  |
| O\_T histogram shows that most of the products were bought from carrefour. |  |
| I\_O\_Sales histogram shows that the most common sales of the product in each store in dollars were between 0-2000 dollars. |  |

* 1. Data preprocessing Description and justification

|  |  |  |  |
| --- | --- | --- | --- |
| **Step** | **data pre-processing step** | **Description** | **Justification** |
|  | Data cleaning: (I\_vis) column | I used the replace function to replace the zeros in the column with null values, and then I filled them with the mean using the fillna() function along with mean() function. | I\_Vis feature represents how much space is dedicated to displaying a specific product within a store therefore it shouldn't contain zeros. |
|  | Data cleaning: (I\_W) column | I used the fillna() function along with mean() function, to fill the missing values in the (I\_W) feature. | Since I\_W feature contains continues values and the feature shouldn’t contain null values to make better prediction. |
|  | Data cleaning: (I\_Id) column | I used the drop () function to drop this feature. | Because it is an unnecessary feature and the KNN and linear regression gave better results when dropping it. |
|  | Data cleaning: ( I\_MRP\_JD ) column | I used the drop () function to drop this feature. | because the correlation between(I\_MRP\_JD)and(I\_MRP\_US) is high. |
|  | Data cleaning: ( O\_Size) column | I used the drop () function to drop this feature. | Because it was related to the O\_Id feature, as there were similar outcomes between them, although it didn’t show a high correlation between them. But it gave better results when dropping it. |
|  | Data integration: (I\_Fat\_C) coulumn. | I used the replace function to replace the duplicated data with different names. | Because there shouldn’t be multiple data that means the same thing in different names as it would give bad results in machine learning. |
|  | Data Transformation using normalization:  (I\_W) column | I used both the min-max normalization as well as the z-score normalization to scale the data within a range. | In order to ensure that the data is formatted and scaled appropriately, as this helps in improving the performance of the machine learning models. |
|  | Data Transformation using normalization:  ( I\_Vis ) column | I used both the min-max normalization as well as the z-score normalization to scale the data within a range. | In order to ensure that the data is formatted and scaled appropriately, as this helps in improving the performance of the machine learning models. |
|  | Data Transformation using normalization:  ( I\_MRP\_US ) column | I used both the min-max normalization as well as the z-score normalization to scale the data within a range. | In order to ensure that the data is formatted and scaled appropriately, as this helps in improving the performance of the machine learning models. |
|  | Data Transformation using encoding:  ( O\_Id ) column | I used get\_dummies to transform this column into numbers. And I dropped the o\_Id column and added the transformed data to the main dataset. | In order to turn the categorical values into numbers to make it easier for the machine learning to deal with, also this column contain nominal variables that’s way I used get\_dummies |
|  | Data Transformation using encoding:  ( I\_category) column | I used get\_dummies to transform this column into numbers. And I dropped the o\_Id column and added the transformed data to the main dataset. | In order to turn the categorical values into numbers to make it easier for the machine learning to deal with, also this column contain nominal variables that’s way I used get\_dummies |
|  | Data Transformation using encoding:  ( O\_Loc\_T ) column | I used fit transform function from label encoder to transform the data into numbers. | In order to turn the categorical values into numbers to make it easier for the machine learning to deal with, also this column contain ordinal variables that’s way I used (LabelEncoder()) |
|  | Data Transformation using encoding:  ( I\_Fat\_C ) column | I used fit transform function from label encoder to transform the data into numbers. | In order to turn the categorical values into numbers to make it easier for the machine learning to deal with, also this column contain ordinal variables that’s way I used (LabelEncoder()) |
|  | Data Transformation using encoding:  ( I\_Recalled ) column | I used fit transform function from label encoder to transform the data into numbers. | In order to turn the categorical values into numbers to make it easier for the machine learning to deal with, also this column contain ordinal variables that’s way I used (LabelEncoder()) |

1. Experiments:
   1. Linear Regression Model-done in code.
   2. Classification Model-done code.
   3. Compare the different models.

In the code I used two machine learning models, the Linear Regression as well as the k-Nearest Neighbors (k-NN) algorithms.

**Linear regression**:

Linear regression is used in order to predict continues values, such as the I\_O\_Sales in the project. Its job is to find the linear relationship that shows the connection between dependent variable and the independent variables.

The goal is to find the least differences between the predicted values and the actual values. Moreover, the linear regression presume that the dependent and independent variables have a linear relationship.

**k-Nearest Neighbors (k-NN):**

I used the KNN algorithm to predict the dependent variable O\_T. The KNN works by looking for the data points that are closer to the new data point by choosing a k value. After that the majority of these existing data points are used to predict the class of the new data point.

In conclusion, the linear regression algorithm is used with continuous variables, and it assumes that the relationship between the independent and dependent variables is a linear relationship. KNN on the other hand, is used for both regression and classification but it does not make assumptions about the data, nevertheless it does assume that the similar points are close together.

* + 1. Linear Regression Model with different learning rates

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Alpha (LR)** | **R2** | **MSE** | **RMSE** | **MAE** |
| **LR=0.1** | -2.829962e+118 | 8.617466e+124 | 2.935552089776661e+62 | 2.935526e+62 |
| **LR=0.01** | -3.963809e+98 | 1.207012e+105 | 3.474208402439544e+52 | 3.474160e+52 |
| **LR=0.001** | -5.517513e+78 | 1.680128e+85 | 4.0989365234940735e+42 | 4.098901e+42 |

* + 1. KNN Model with different k values

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **K value** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| **K=3** | 0.76 | **0.74** | 0.76 | **0.74** |
| **K=5** | 0.77 | 0.73 | 0.77 | **0.74** |
| **K=7** | **0.78** | **0.74** | **0.78** | 0.73 |

* + 1. Linear Regression Model with normalization

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Approach** | **R2** | **MSE** | **RMSE** | **MAE** |
| without normalization | -2.829962e+118 | **8.617466e+124** | 2.935552089776661e+62 | 2.935526e+62 |
| Min-max normalization | -8.634916e+118 | 2.629403e+125 | 5.127770530657142e+62 | 5.127726e+62 |
| Z score Normalization | **-1.175561e+119** | 3.579680e+125 | **5.983042519297787e+62** | **5.982990e+62** |

* + 1. KNN Model with normalization

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Approach** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| without normalization | 0.76 | 0.74 | 0.76 | 0.74 |
| Min-max normalization | **0.81** | **0.80** | **0.81** | **0.80** |
| Z score Normalization | **0.81** | **0.80** | **0.81** | **0.80** |

* 1. Analysis of the results

**Linear regression with different learning rates:**

In my project I used three different learning rates. 0.1,0.01, and 0.001. the **R2** results were getting smaller when the learning rate got smaller. As it began with

**-2.829962e+118** when the learning rate was **0.1**, and then **-3.963809e+98** with a learning rate of **0.01**, and the least one was **-5.517513e+78** with a learning rate of **0.001**. the same was for the **MSE** as it began with **8.617466e+124** when the learning rate was **0.1** and reduced to around **1.5** when the leaning rates were **0.01** and **0.001**. on the other hand, the results of **RMSE** and **MAE** were getting higher when reducing the number of the learning rate. As they both started with around **2.9** at **0.1** learning rate to reach almost **4.1** at **0.001** learning rate.

**Linear regression and data normalization:**

In my project I used three approaches for the linear regression model. I tried training the model at first without using normalization and then I tried it with both the min-max normalization and the z-score normalization. In general, the **R2, RMSE**, and the **MAE** showed the highest results using the Z score normalization. Although, **MSE** showed the highest result without normalization. The **R2** had the lowest result of approximately **-8.7** with the min-max normalization, and the highest result of almost **-1.2** using the Z score normalization. The **MSE** had the lowest result of **2.6** with min-max normalization, and the highest result of **8.6** without normalization. The **RMSE** and the **MAE** least result was almost **3.0** without normalization, and the highest result of approximately **6.0** with the Z score normalization.

**k-Nearest Neighbors (k-NN) with different K values:**

In my project I used three different K values. K=3, K=5, and K=7. Overall the best results were when K was equal to 7. Starting with the **accuracy**, the least result was when **K=3** as it had an accuracy of **0.76**. and the best result was when **k=7** as the accuracy was **0.78**. moving on to the **precision**, it had the highest result of **0.74** when **k=3** and **k=7**, and the least result was **0.73** when **k=5** which wasn’t that much of a difference. The **recall** was the same as the accuracy, as the least result was when **K=3** as it had a recall **of 0.76.** and the best result was when **k=7** as the recall was **0.78**. the **F1 score** was the only one different, as it’s highest result’s were with **k=3**, and **k=5** as it had a high score of **0.74**, and the least was **0.73** when **k** was **7**.

**k-Nearest Neighbors (k-NN) and Normalization:**

In my project I tired three techniques. I trained the model at first without normalizing the data and the I tried the min-max and Z score normalization.

The Z score and the min-max normalization had the best results in all of the evaluation measures. Although, it got the worst results when I didn’t normalize the data.

Both the **accuracy** and **Recall** got **0.81** in both normalization methods. In addition, the precision and **F1** score got **0.80** in both normalization methods. The worst results were **0.76** for the accuracy and recall and **0.74** for the **precision and F1** score when I didn’t normalize the data.

* 1. The effectiveness of different models

Starting with the KNN, the effectiveness of the KNN depends on the value of K. as when the value of K becomes larger the model is supposed to become less sensitive to any noise in the data. on the other hand, a smaller value of K makes the model more sensitive to the noise in the data. In my project the experiment that gave the best results were when I used K=7 as it gave the best results in all the evaluation measures.

The linear regression on the other hand, depends on the linear relationship between dependent and independent variables. As long as the relationship is strong, the model of linear regression is supposed to perform really well. Also, another important thing to consider is the quality of the data, such as if it contains outliers or it was not well cleaned, this well affect the performance model.

I prepare using KNN as it is a simple and more understandable way, as well as it one of the best training techniques. The great thing about KNN is that it takes the training data and stores it and waits until a new data point in order to make predictions by finding the nearest neighbors of the new data point, and then use the class that occurs the most between all of them to make the prediction.

1. Libraries used in the project:

Python contains many libraries that gives capabilities that are better than the language’s built-in capabilities. They are a pre-Witten codes that are imported easily. they could be used for multiple tasks, such as manipulating data, and machine learning.

Some of the most important libraries are (NumPy, pandas, scikit learn):

1. NumPy:

* NumPy is a library from python that is used with multidimensional arrays as well as matrices.
* It contains a lot of tools in order to perform mathematical operations for matrices and arrays, which make it easier in numerical as well as scientific computations.
* The NumPy arrays are used to store big amount of data and make quick operations on them.

1. Pandas:

* Pandas’ library is very important for handling data and manipulation through giving simple data structures in addition to better capabilites for data analysis.
* Pandas are used to ease the work with data that are structured.
* It is used with data science and projects for machine learning.
* It is important in data visualization and data cleaning.

1. Scikit learn:

* It is a library that is mainly used for machine learning, as it gives ability to perform supervised as well as unsupervised learning models.
* It contains algorithms for regression, clustering, pre-processing, and classification.

1. Matplotlib:

* It is one of the most important libraries in data visualization.
* It gives a lot of ways to visualize your data, such as the bar charts, radar charts, line plots.

|  |  |  |
| --- | --- | --- |
| **Student name:** Yousef Abu Ali  **Student ID: 21110294**  **Is the student repeating this unit? YES NO** | | **Assessor name:**  Dr. Murad A. Yaghi  Dr. Heba Alawneh  Bassam Kasasbeh |
| **Issue date:**  08-01-2023 | **Submission date:**  30-01-2023 | **Submitted on:**  **30/01/2023** |
| **Programme:** Computing | | |
| **Course Name: Principles of Data Science and Computing Systems**  **HTU Course Code: 10204280 BTEC UNIT:** | | |
| **Assignment number and title: 2 / Application of Data Science Life Cycle** | | |

**Plagiarism**

Plagiarism is a particular form of cheating. Plagiarism must be avoided at all costs and students who break the rules, however innocently, may be penalised. It is your responsibility to ensure that you understand correct referencing practices. As a university level student, you are expected to use appropriate references throughout and keep carefully detailed notes of all your sources of materials for material you have used in your work, including any material downloaded from the Internet. Please consult the relevant unit lecturer or your course tutor if you need any further advice.

|  |
| --- |
| **Student declaration**  I certify that the assignment submission is entirely my own work and I fully understand the consequences of plagiarism. I understand that making a false declaration is a form of malpractice.  **Student signature: Date: 30/01/2023** |

References:

<https://www.geeksforgeeks.org/supervised-unsupervised-learning/>

<https://www.geeksforgeeks.org/ml-classification-vs-regression/>

<https://www.statology.org/regression-vs-classification/>

<https://en.wikipedia.org/wiki/Amazon_Web_Services>

<https://aws.amazon.com/about-aws/>

<https://www.simplilearn.com/tutorials/azure-tutorial/what-is-azure>

<https://www.techtarget.com/searchcloudcomputing/definition/Windows-Azure>