IE30301-Datamining Assignment 2

Assignment Description

There are two types of problem sets for this Assignment2: **Theoretical Derivations(70 points) + Programming Assignments(30 points)**. The description for each Assignment is below. Please read the instruction carefully. If you have questions regarding the homework, please use the discussion board. TA will not reply to emails regarding assignments to avoid any possible information disparity between students. Please submit your Assignment to the blackboard.

- ✓ **Due Date:** 2022.03.30(Wednesday) 12 pm Korea Standard Time.
- ✓ **We will not accept late work!!**
- ✓ **Handwriting is only allowed for Exercise 3.1 and 3.2 in TheoreticalHW. You may take a photo or write in digital tools(Galaxy Tap, Ipad, etc.). Paste your solution in your submission file.**
- ✓ **Submission format:** Please submit a .zip file containing the below **four** files. Follow the naming rules as stated.

your_student_ID_Assignment2_your NAME.zip

your_student_ID_ProgrammingHW2_yourNAME.pdf

your_student_ID_TheoreticalHW2_yourNAME.pdf

your_student_ID_TheoreticalHW2_ORIGINAL_yourNAME.tex

your_student_ID_ProgrammingHW2_yourNAME.ipynb

Ex. 20215318_ProgrammingHW2_HongGilDong.pdf

Collaboration Policy & Academic Integrity

- ✓ Study groups are allowed, and students may discuss in groups. However, you cannot discuss or share solutions openly through Kakao open group chats, Chegg, or any other places where it is not legitimate. The discussion should only be "high level" and "general." We expect students to understand and complete their assignments. Each student must write down the solutions independently.
- ✓ It is not allowed to get solutions from other students.
- ✓ If you referred to specific webpages or exterior materials, please write down the link on the Assignment.
- ✓ If you worked in a group, please put the names of your study group on your Assignment on top
- ✓ You are not allowed to contact the TA for help personally. Even if it is a trivial question, we are not allowed to answer. Please use the discussion board.
- ✓ HONOR CODE: we take the UNIST student HONOR CODE seriously. Any form of cheating and plagiarism is not allowed.

Theoretical Derivations

- **Directions :**
  - Solve the questions in 'IE30301_Assignment2.pdf.'
  - **Exercise 1,2,4,5, 3.3, 3.4** -> We will only accept typed-in latex file format, which means that we will not take handwritings. Files that are written on tablets (Galaxy Tab, Ipad, etc.) will not be accepted
  - **Execrise 3.1, 3.2** -> We accept both handwriting and typed in files. You can write your results on a paper, take a photo and paste it on the **'Homework2_template.tex' by using *LaTeX* commands**. You may use any digital writing solutions such as Ipad or Galaxy Tab. Please write your writing neatly and use a high resolution photo image. **If the writing is illegible or the photo is not recognizable, we will not be able to give points.**
  - Do not redistribute the paper or upload the questions online.
  - Please use *LaTeX* with overleaf to type in all your answers. Convert your file to pdf format for submission.
  - We will provide you file 'homework2_template.tex' as a template for the answer sheet. Please be mindful that we will only accept the answer sheet which is written based on this template file. You should submit tex file from this template and pdf file converted from that tex file.
  - There are a total of **5 exercises,** each with different points
  - copying code or solution → NO MERCY **F**
  - Note that only 50% of the questions will be randomly selected and graded.

Programming Assignments

- **Directions :**
  - Use the attached 'IE30301_Assignment2_Skeleton.ipynb' Jupyter Notebook file and fill each cell through the numbered tasks.
  - You must use '*Python*' as a programming language.
  - You can use only basic libraries (e.g., NumPy, Pandas) and designated packages, including :
    (* You don't have to use all the below packages. Just choose what you want among the packages)
    - Matplotlib
    - Scikit-learn
    - Statsmodels
  - Both *Scikit-learn* and *Statsmodels* will allow you to implement a linear regression algorithm, while deriving the p-value of each independent variable requires you to use *Statsmodels*.
  - Convert the completed 'ipynb file' to 'pdf file' for the final submission.
    - Submit **both 'ipynb file' and 'pdf file'**

[Task 1] Multiple linear regression

1.1 Load dataset.

- Read the 'diabetes.csv' file into a DataFrame.

1.2 Split Dataset into the train & test set.

- You have to use the specified random seed just before dividing the Dataset using the package function, 'train_test_split'. (The python seed method is used to initialize the pseudorandom number generator. If you provide the same seed value before generating random data, it will produce the same data)
- The specified **random seed value is 0.**
- The train set and test set ratios are as follows: 70% train set / 30% test set.

1.3 Preprocess the dataset using a standard scaling method.

- You have to use the standard scaling method for preprocessing the dataset and for that, specific function, 'StandardScaler' of scikit-learn should be used.

1.4 Import a Linear Regression algorithm from a package you choose and train your model.

- When you train your model, you must include an **intercept**. Using the package algorithm's default setting can cause you not to get an intercept value for your model. Consider a way to deploy your model with an intercept term by considering another function of packages.

1.5 Predict the values of the test dataset's target variables through your trained model and evaluate the model performance meaning how much gap between true values and predicted values has occurred on your test dataset.

- Predict the target variable values by using the trained model.
- Evaluate your trained model's performance using the following two metrics (***It is NOT ALLOWED to use packages for this task. You must compute two metrics by your own code without ready-made packages**.)
  - mean squared error (MSE)
  - R2 score

1.6 Write down the linear equation formula of the trained model by *LaTeX* style in the notebook. (Referring to the site [URL])

1.7 Check the p-value of independent variables of your model and find out significant variables.

- Print the p-value of each variable.

- List all variables judged to be significant based on the p-value of 0.05.

[Task 2] Simple linear regression

2.1 Select a single feature using the information of p-value.

- Select one of the significant variables in **'task 1.7'** and print the selected independent variable name. There might be several essential variables in terms of the p-value. You can choose one of them for building a simple linear regression.

2.2 Split Dataset into the train & test set.

- You have to use the specified random seed when dividing the Dataset using the package 'train_test_split'.
- The specified random seed value is 0.
- The train set and test set ratios are as follows: 70% train set / 30% test set.

2.3 Preprocess the dataset using a standard scaling method.

- You have to use the standard scaling method for preprocessing the dataset and for that, specific function, 'StandardScaler' of scikit-learn should be used.

2.4 Import Linear Regression algorithm from a package you choose and train your model.

- When you train your model, you must include an intercept. Using the package algorithm's default setting can cause you not to get an intercept value for your model. Consider a way to deploy your model with an intercept term by considering another function of packages.

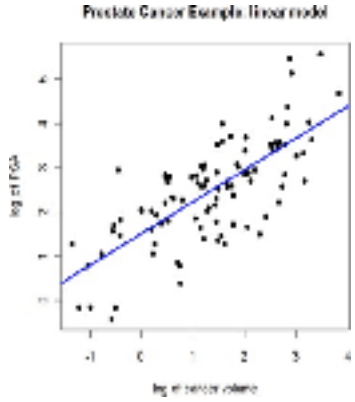
2.5 Predict the values of the test dataset's target variables through your trained model and evaluate the model. performance meaning how much gap between true values and predicted values has occurred on your test dataset.

- Predict the target variable using the trained model.
- Evaluate your trained model's performance using the two metrics (*it is not allowed to use packages for this task. You must compute two metrics by your own code without ready-made packages.)
  - mean squared error(MSE)
  - R2 score

2.6 Write down the linear equation formula of the trained model by *LaTeX* style in the notebook.

2.7 Plot a regression line found by training dataset. Plot the test data points as well.

- Plot data points of the test set and draw the line derived from the trained model as below picture.
- Describe the title of the graph and the names of the two axes as below picture.


Prostate Cancer Example: linear model
log of PSA
log of cancer volume
-1
0
1
2
3
4