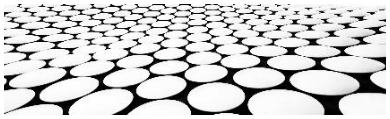
CSI4142 Fundamentals of Data Science



ASSIGNMENT 3

Predictive analysis
Regression and Classification

GOALS

The overall goal of this assignment is to become familiar with Predictive Analysis, both for regression (prediction of numerical values) and classification (prediction of a categorical value).

At the end of this assignment, you will have:

- Reviewed your Python skills, as the project MUST be done in Python
- Explored Kaggle, an amazing resource of challenges and datasets
- Programmed and evaluated a linear regression approach on a regression task
- Programmed and evaluated the impact of outlier removal on a linear regression method
- Programmed and evaluated a decision tree approach on a classification task
- Programmed and evaluated the impact of feature aggregation on both linear regression and decision trees
- Documented everything, in Jupyter Notebooks, in a way to make your analysis understandable and reproducible.

SUBMISSION INFORMATION

- Deadline:
  - Submission of your notebook: **Tuesday, March 11th, midnight**
  - Grace period (no penalty) is until 12:30… after that it's 20% per day off.
- Groups:
  - You are expected to form groups of 2 and do a single submission per group. Please use the same groups as for Assignment 1&2. If for some reason, you need to modify your group, send me an email as some manual changes will need to be done in Brightspace.
  - If you prefer to work alone, that is accepted (but not encouraged), but the requirements are not changed.
- Where to submit:
  - Your submission must be done in Brightspace in Assignment section (Assignment 3)
- Submission format:
  - Only Jupyter Notebook files will be accepted to show your programs
  - (optional) You are encouraged to submit a PDF or HTML version of your notebook as supplementary evidence of your work.
  - For the datasets used in your analysis, your code MUST contain a link to the dataset so the TAS do NOT need to download any data.

You should submit 2 Jupyter Notebook, one for regression (Dataset 1) and one for classification (Dataset 2).

Each Jupyter Notebook should include:

1. Group number, names and student numbers of group members
2. Introduction (what is the notebook about, and how to use it)
3. Dataset description (Dataset 1 or Dataset 2)
4. Regression or Classification Empirical study (Steps (a) to (g) described in Sections 2 and 3)
5. Conclusion (one paragraph summary of what was achieved and what more could be done)
6. References

PLEASE NOTE:

If the corrector cannot run your code, the mark will be zero for the entire assignment. It is your responsibility to test that the code cells in your Notebook are executable. The most frequent problem is that the dataset is not readable (file not found). Please make sure the dataset is either read directly from a public place (a dataset repository) or read from one of your shared repository (accessible to anyone) so the TA can run your notebook without any data download.

PROGRAMMING REQUIREMENTS

1. Choose 2 datasets

As you know, the Kaggle site is a very good site to explore as it contains datasets for many tasks in data science and AI. There are also other sites for datasets if you wish to further explore. You must choose 2 datasets, and they cannot be the same as for your Assignment 1 nor Assignment 2.

Dataset 1 – Regression

The first dataset will serve for your empirical study of regression (described in Section 2).

For **Dataset 1**, you must choose among the 4 datasets suggested below by the TAs. If, by any chance, you have already worked with one of the datasets below in Assignment 1 or 2, please do not select it. The TAs mentioned that the following datasets contain possible outliers and categorical data, making them ideal for the first part of the assignment.

Suggested Datasets

1. **Car Price Prediction Dataset**
   - **Description**: Predicts car prices based on features like mileage, brand, year, fuel type, and transmission type.
   - **Categorical Features**: Car brand, fuel type, transmission type.
   - **Outliers**: Potential outliers in price and mileage.
   - **Link**: Kaggle Car Price Dataset
2. **Medical Insurance Cost Prediction Dataset**
   - **Description**: Predicts medical insurance costs based on age, BMI, smoking status, region, and other features.
   - **Categorical Features**: Smoking status, region, gender.
   - **Outliers**: Potential outliers in insurance costs.
   - **Link**: Kaggle Medical Insurance Dataset
3. **Housing Prices Regression Dataset**
   - **Description**: Predicts house prices with a couple of binary features.
   - **Categorical Features**: Binary features (e.g., presence of a garage, basement, etc.).
   - **Outliers**: Clean dataset, but outliers can be simulated if needed.
   - **Link**: Housing Prices Regression
4. **Laptop Price Prediction Dataset**
   - **Description**: Predicts laptop prices based on specifications like brand, RAM, storage, and processor.
   - **Categorical Features**: Brand, processor type, operating system.
   - **Outliers**: Potential outliers in price or specifications.
   - **Link**: Laptop Price Prediction

Dataset 2 – Classification

The second dataset will serve for your empirical study of classification (described in Section 3).

For **Dataset 2**, you must select a dataset that is NOT within the 4 suggested above. In fact, the previous datasets are for regression analysis and you must find a dataset for classification analysis, so they would not qualify.

Furthermore, your second dataset must be in a domain different from the first dataset. I want you to explore 2 different domains. Another constraint is to find a dataset with a minimum of 5 columns so you have various features to explore.

Please note: During the semester, for your assignments, you will need to choose various datasets for different explorations so you are exposed to data from different domains. You are NOT allowed to reuse a dataset from one assignment to another. So, if you find a few datasets that you think are interesting, save them for the last assignment.

2. Regression Empirical Study with Linear Regression

In this part of the assignment, you must perform an empirical study in which you evaluate a linear regression approach on a regression task.

a) Cleaning the data
   a. You have programmed in Assignment 2 some validity checks and imputation methods. Reuse your methods here to clean your data.

b) Categorical feature encoding
   a. Linear regression applies to numerical features. We have seen in class that one-hot encoding can transform categorical features into numerical features. Apply such encoding.

c) EDA and Outlier detection
   a. We explored in class different approaches for outlier detection, one of them being Local Outlier Factor (LOF). Program such approach for outlier detection.
   b. As LOF is costly to use (if large dataset), first use EDA to visualize the data and find which feature LOF should be used on. If there are no features with outliers in your dataset, then purposely introduce some outliers for at least one feature so you can use LOF to detect them.
   c. Decide what to do with outliers. Either remove them, or consider them as missing values on which to apply imputation methods.

d) Predictive analysis: Linear regression
   a. Explore the *LinearRegression* method suggested in scikit-learn (or other packages). scikit-learn also contains Ridge, Lasso and ElasticNet regressions if you want to explore those (optional).

e) Feature Engineering

   a. Program a feature aggregator to create 2 additional features. We will explore aggregation during March 5th lecture, but you can go ahead earlier and program something simple (e.g. adding two features, ratio, count). You can also try to introduce squared feature, or multiply different features together. Think of interesting ways to combine the features based on the dataset and what could make sense. The purpose of aggregation is to create new features from the ones you have.

f) Empirical study

   a. Split your data into 3 subsets for train, validation and test sets.
   b. Decide on some evaluation metric(s) (MSE, LMSE, R2). Choose at least one.
   c. As baseline, use the linear regression with baseline settings, without outlier removal and without feature aggregation. Using your chosen metric(s), evaluate on the validation set using a 4-fold cross-validation.
   d. Try different combinations of with/without dealing with outliers, with/without feature aggregation. Using your chosen metric(s), evaluate the different system versions on the validation set using a 4-fold cross-validation. Based on those experiments, you will choose the best system.
   e. Decide on a FINAL system, which is the best one according to your empirical study and use the Test Set to perform your final evaluation. The Test Set should NOT have been used before this final evaluation.

g) Result analysis

   a. Analyse the obtained results and show if any improvement was achieved with different settings.
   b. Discuss the impact of outlier detection and feature aggregation.
   c. Discuss how the results on the unseen test set compare to the cross-validation results.

3. <u>Classification Empirical Study with Decision Trees</u>

In this part of the assignment, you must perform an empirical study in which you evaluate a decision tree approach on a classification task.

a) Cleaning the data

   a. (same as for regression study)

b) Numerical feature encoding (optional)

   a. Decision trees might benefit from binning of numerical values. You can explore this idea, but it's not mandatory.

c) EDA and Outlier detection

   a. (same as for regression study)

d) Predictive analysis: Decision Trees

   a. Explore the *DecisionTreeClassifier* method suggested in scikit-learn (or other packages).
   b. Look at the parameters (splitting criterion (gini, entropy), max_depth, min_samples_split, etc) and choose a baseline setting.

e) Feature Engineering

   a. *(same as for regression study)*

f) Empirical study

   a. Split your data into 3 subsets for train, validation and test sets.
   b. Decide on some evaluation metric(s) (Precision, Recall, F1, Accuracy). Choose at least one.
   c. As baseline, use the *DecisionTreeClassifier* method with baseline settings, without outlier removal and without feature aggregation. Using your chosen metric, evaluate on the validation set using a 4-fold cross-validation.
   d. Try different combinations of with/without dealing with outliers, with/without feature aggregation. Using your chosen metric(s), evaluate each system version on the validation set using a 4-fold cross-validation. Based on those experiments, you will choose the best system.
   e. (optional) You can also try different parameter settings for the *DecisionTreeClassifier* method (changing the criterion, max_depth, min_samples_split) and use cross-validation evaluation to decide on a best setting.
   f. Decide on a FINAL system, which is the best one according to your empirical study and use the Test Set to perform your final evaluation. The Test Set should NOT have been used before this final evaluation.

g) Result analysis

   a. Analyse the obtained results and show if any improvement was achieved with different settings.
   b. Discuss the impact of outlier detection and feature aggregation.
   c. Discuss how the results on the unseen test set compare to the cross-validation results.

PLEASE NOTE: Data cleaning, outlier detection and feature engineering are part of both studies. You can reuse the same code.

☆☆☆

EVALUATION (50 points)

- Overall effort in the reports (5 points)
  - Provided all required steps clearly identified with different sections
  - Good cell separation (text, code, results, etc)
  - Tests on various examples easy to perform by the corrector
  - Report detailed enough for reproducibility
- Datasets descriptions (4 points)
  - Dataset name, author, purpose (what was it made for)
  - Shape: how many rows and columns
  - A list of the features + descriptions (what do they represent and are they categorical or numerical)
- Data Cleaning, Outlier Detection, feature aggregation (10 points)
  - Data cleaning tools used on both datasets
  - Programming of outlier Detection with LOF and decision on how to deal with outliers for both datasets
  - Programming of feature aggregation to create at least 2 additional features on both datasets
- Linear Regression study (14 points)
  - Correct use of train/validation/test
  - Correct use of metrics
  - Correct experiments
  - Clear presentation of results (comparative table + analysis)
- Classification study (14 points)
  - Correct use of train/validation/test
  - Correct use of metrics
  - Correct experiments
  - Clear presentation of results (comparative table + analysis)
- References (3 points)
  - For any part of your code taken from a web site (even a tutorial site or stackoverflow), you must provide the reference to it.
  - If part of your code was generated using Generative AI, please list the tool as well as the queries performed for each code snippet used.
  - Any theory/algorithms found in books, slides, tutorials that you used should be referenced.

QUESTIONS

- You can ask your questions within the assignment topic of the discussion forum on Brightspace.
- To make sure you get answers on the forum on a regular basis, the 3 TAs will share the days to answer your question. Notice that they are NOT available on weekends.

	Monday	Tuesday	Wednesday	Thursday	Friday
Morning	Gurdarshan	Gurdarshan	Bhavneet		Ángel
Evening	Ángel	Gurdarshan	Ángel	Bhavneet	Bhavneet