BLG 454E - Learning from Data
Fall – 2022
-Term Project-

Problem: Prediction of insurance costs.

Deadline: 19 December 2022, 11:59 PM

The term project has 3 deliverables: The Kaggle competition results, the source code and a report explaining the designed framework and presenting your results. You will be provided with **templates to use** for the report.

1. Kaggle Competition (20 points)

Description

We have created a private class competition on Kaggle. Please click the following link for the term project competition:

https://www.kaggle.com/t/2c268433178d4d42a7dedc40f9a3ce56

Dataset

The power of data can be leveraged to predict the unknown. One of the most prominent application domains is price prediction. Predicting the price of anything provides us with valuable information and insight. Data analytics is especially important in the prediction of insurance fees for both the customer and insurer end. It can help clients make the best decision amongst multiple options and can help insurers determine people appropriate to insure. The price of insurance depends on many different factors. Some of these factors can be listed as age, gender, BMI, and other information related to lifestyle such as smoking.

You are given a dataset of information regarding medical cost. The data consists of the following columns:

- **Age (integer):** age of primary beneficiary
- **Sex (nominal (binary)):** insurance contractor gender, female, male
- **BMI (decimal):** Body mass index provides an understanding of body. It is the ratio of weight to height squared (kg/m^2) and is ideally in the range 18.5 to 24.9
- **Children (integer):** Number of children or dependents covered by health insurance
- **Smoker (nominal (binary)):** Smoking status of the individual
- **Region (nominal):** the beneficiary's residential area in the US, northeast, southeast, southwest, northwest
- **Charges (decimal):** Individual medical costs billed by health insurance

Goal

In this challenge, we ask you to apply the tools of machine learning to predict the cost that an individual must pay for insurance based on the factors listed above.

Submission Process

To see the performance of your model on test data, submit your predictions of test data to Kaggle in the defined format. Kaggle will calculate and rank the submission scores using the public test data throughout the competition. These scores are publicly visible on public leaderboard. After the competition end, a *private* test data is used to calculate final model performance. Private leaderboard is not released to users until the competition has been closed. Public leaderboard is

calculated with 50% of the test data. The final results will be based on the other 50%, so the final standings may be different. Therefore, train your model as general as possible to avoid overfitting on train and public part of the test data.

Scoring Metric

In Kaggle, your submission is evaluated by the Mean Squared Error (MSE).

Submission File Format

You should submit a csv file with exactly [438] x [2] entries plus a header row. The file should have exactly 2 columns:

1. ID: [0,…,437]
2. Predicted (contains your real-valued values)

Submission CSVs must have a header row consisting of ID and Predicted as in the sample submission. Using different column names causes a fail in submission process. ID column must include all ID values between [0, 437].

PS: Your submission will raise an error in cases: you have extra columns (beyond ID and Predicted), extra rows, ID column doesn't consist of integers between [0,437], Predicted column includes values other than real-values.

```
ID,Predicted
0,4232.2231
1,2395.17155
2,10797.3362
3,4149.736
...
436,6313.759
437,8059.6791
```

You can download the sample submission file (sampleSubmission.csv) to have a better idea.

Rules

- Every student has to create a Kaggle account
- Form a team of **3 to 5 students (The "team" tab on the competition)**
- Individual submissions are **not allowed**. In such a case, send us an email so that we assign a random teammate.
- Team members **must** be students **officially registered** to the LFD class
- Team names should be in the following format: StudentID1 _StudentID2_StudentID3
- Submission format is explained and a sampleSubmission file (sampleSubmission.csv) is given in the competition webpage.
- You are allowed to use only **Python** programming languages (with jupyter) for the implementation.
- You are only allowed 10 submissions per **DAY**. Start early so you can submit more submissions.
- Academic dishonesty including cheating, plagiarism, and direct copying is unacceptable. Note that your codes and reports will be checked using plagiarism tools!

2. Report (40 points)

Prepare a report in Latex/Word using provided IEEE Conference Paper template. Your report must **not exceed** 2 pages (**one extra page** can be allowed for the **main Figure** illustrating the learning pipeline)!

The report should consist of the following sections:

1. (**6 points) Introduction:** Mention about what and why you did in this project briefly. Give your final score and rank in the competition with your Kaggle name and team name.

2. **(6 points) Datasets:** Explain your methods for data preprocessing in detail.

3. **(20 points) Methods: The how?** Describe each component of your regressor. Include a **main figure** illustrating the key steps of the proposed solution (learning pipeline). Explain how you train and test your model in general. **The why?** Explain **why** you have made selected such components. Give all details about the methods like the algorithms used, parameter tuning, etc.

4. **(6 points) Results and Conclusions: First**, report your **5-fold cross-validation** results. Explain your results. **Second**, give your Kaggle score and ranking.

5. **(2 points) References:** The list of references cited in the report. Don't forget the citation to the related reference in the report.

3. Code with 5-fold CV (30 points)

The version of your code that you will upload should have 5-fold cross-validation implemented. The code should take the train features as input, and perform 5-fold cross-validation for training and validating the designed framework model. The code will have two outputs: (1) the predicted insurance costs saved in a predictions.csv file (you can use the same Kaggle format to same them), and (2) the MSE between the ground truth and predicted samples.

Important note 1: the code will take in 900 samples provided to you and perform 5-fold CV on this set. At this stage, you don't need to use the extra test set that's used in evaluation for Kaggle competition

Tidy up your code as to

- run simply,
- get all necessary inputs as function parameters (train and test data, model parameters),
- produce output, i.e. the submission file (test predictions)
- have explanatory comments

Important note 2: Use the following random anchorization seed when applying 5-fold CV:

```
- import random as r
- r.seed(1)
```

Important note 3: For computing the MSE, once you complete your 5-fold CV, you will end up with predicted vectors and ground truth (actual) vectors for the 900 samples. You can compute the MSE between their vectorized versions as follows:

```
- from sklearn.metrics import mean_squared_error as mse
- mse(predicted,actual)#returns mse result for two melted matrices
```

Important note 4: You can use the pandas module to read in your CSV files

(https://pandas.pydata.org/docs/reference/api/pandas.read_csv.html)

```
- import pandas
- pandas.read_csv([ARGS])# Read a csv files into DataFrames.
```

4. Project Overall Evaluation

For the project, you will provide a final report in IEEE conference paper format (that is given to you in both Word and Latex format). Total score of your project will be calculated as follows:

- The Kaggle competition (50 points)
  - 30 points: 5-fold CV
  - 20 points: Your Kaggle rank
- Report: 40 points
- Code: 10 points
  - Your code should be clean and readable. Implement your code as powerful as possible befitting for a 4th grade student. Weak coding might cause losing 5 to 10 points.

Bonus Marks

Top five teams will be rewarded with bonus marks, 30pts, 25pts, 20pts, 15pts and 5pts. respectively, according to the average of the public and private leaderboard scores.

Ninova Submission Policy

- Submit your PDF report, and code in a zip file through Ninova on time.
    - **Unnecessary uploadings (files, pictures, etc.) will be penalized!**
    - **Only put things in your zip file that you are asked to.**
- No late submissions will be accepted.

References

To learn more about Kaggle Competitions, https://www.kaggle.com/docs/competitions

---

Res. Asst. Doğay KAMAR, kamard@itu.edu.tr

Res. Asst. Meral KUYUCU, korkmazmer@itu.edu.tr