Module

PGR304- Predictive Analytics

Due date for submission

(see Wiseflow)

Subject's supervisor

Noha El-Ganainy

Teacher and e-mail

Vahid Hassani | vahid.hassani@oslomet.no

Learning outcomes

After successfully completing the course the student:

Knowledge

The student...

- can explain the statistical concepts and techniques (such as regression, time series, VAR models and Logit models) for predictive analytics
- knows the statistical concepts underlying the design and analysis of predictive analytics techniques appropriate for a given data science problem
- knows the predictive analytics methods and tools for various data science domain

Skills

The student...

- can select appropriate predictive analytics methods and tools (such as regression, time series, VAR models and Logit models) for a given data science problem
- can analyse empirically the performance of predictive analytics methods and techniques
- can transfer learnings from the course to new data science problems in terms of selection of appropriate predictive analytics methods, techniques and tools
- can use python or another natural language processing platform to implement predictive analytics solutions

General competence

The student...

- can discuss concepts and applications of predictive analytics with peers
- can differentiate the suitability and efficiency of programs in terms of the predictive analytics methods and techniques employed
- can apply the knowledge of and skills in predictive analytics in various data science domains
- can critically reflect on the tradeoffs in the design and implementation of predictive analytics solutions

Please address the following questions (cases) in your submission.

This assignment must be solved individually and counts for 100% of the total grade. The weighting of each sub-task is stated as a percentage.

You have been given access to this dataset and must first spend some time getting to know it.

The dataset refers to an experiment in a factory. The owner has been running 480 separate experiments, where the factory produces its product under different setting. (for example see this as production line for an specific chocolate bar)

The owner of the factory has reported 25 different values for each experiment.
The Features1 to Feature22 , are some statistical calculations for some variables that are measured in the factory during each experiment (like mean value, Standard deviation, minimum or maximum value of some specific sensors during the length of the experiment, and so; these sensors are measuring different variables like, pressure, density, etc).

The Feature23 and Feature24 are two input variables that can affect the results of each experiment (consider something like a temperature of the factory; this is just an example). The Feature23 refers to the start of the experiment and Feature24 is more connected towards the end of the experiment (again if we use the example of the temperature in the factory, then the Feature23 shows the temperature in the start of the experiment and Feature24 shows the temperature close to the end of the experiment).

The last feature is the target variable (final result of the experiment). The target variable is an integer value between 1 and 5. For example you can consider the target variable as an indicator describing the quality of the product that the factory produces (1 referring to the best quality, and 5 meaning the worst quality).

The dataset is in the attached file “DataSet_4_Exam.zip” or “DataSet_4_Exam.csv”. Use Python packages such as Pandas to open the file.

Task 1 - Understanding the data (5%)

- What kind of dataset is this? Provide some context by saying a little about what kind of objects we are talking about; are they correlated or not, ...

Task 2 - Utility value (5%)

- What can you use this data set for? Name at least 2 different applications, or examples of getting value out of the dataset.
- Is the data set ready to be analyzed as it is? Is some data processing necessary before using it? Provide arguments that justify your answer and proceed with your suggestions.

Task 3 – Analysis, modeling and prediction (60%)

The owner of the factory is interested in being able to predict the result of the process based on the either input variable (i.e. Feature23 and Feature24), or all the first 24 Features.

1. What methods would you apply to analyze the data, for the prediction that the owner is interested in (Prediction of the Results of each experiment)? Consider the potential application of machine learning, statistical methods, visualizations, etc. Justify your choices.
2. Partition the dataset into two smaller sets: one small set for future validation (for example 10-15 rows/experiments) and one bigger set for training purpose.
3. Apply Logit models for classification of the results. After tuning the logit model, apply the model to the small set that you kept for validation (and not been used in the training/tuning). Explain the results.
4. Try to reduce the number of the features. Use PCA for doing so. After reducing the dimension of the dataset, repeat the last two subtasks (partitioning the dataset for training and validation and then applying a Logit model)
5. Carry out the analysis suggested above, clearly explaining each step of the procedure you choose to follow.

Bonus Task - Analysis, modeling and prediction (20%)

You can use other tools that you know from your study in other modules or the methods that you have learned yourself. Anything that can be applied for predictive analytics (using the same dataset for this exam). For example, decision trees can be easily applied here. Similarly, one can try KNN or other tools.

Or for reducing the dimension, you could apply Linear Discriminant Analysis (LDA) instead of PCA or other approces.

- In case you want to answer to the bonus Task: Repeat the whole **task 3** with the tools/methods that you have learned on yourself or from other courses.

Task 4 - Results and evaluation (30%)

- What did you get out of the data? Show concrete numbers, figures and graphs.
- Carry out a critical reflection on the result and the insight you have arrived at. Discuss your assumptions and the range of validity of your predictions in detail.

Assignment specification

1. A report (3500 - 5000 words) that answers all the assignment questions. Remember to include illustrations where appropriate. **The word count is excluding the phyton codes**.
2. The report must also contain a separate section that shows an overview of which tools/methods you have used in the assignment.
3. All source code you have written must be included as an attachment to the report (in *.ipynb format).

Assignment criteria*

Grade	Learning Outcome 1: Knowledge	Learning Outcome 2: Skills	Learning Outcome 3: Competence
A Excellent	Excellent and comprehensive understanding of concepts	Demonstrates excellent analytical, technical and writing skills	Outstanding degree of judgment and independent critical thinking
B Very good	Very good understanding of concepts	Demonstrates very good analytical, technical and writing skills	Sound degree of judgment and independent critical thinking
C Good	Good understanding of theory in most important areas	Demonstrates good analytical, technical and writing skills	Reasonable degree of judgment and independent critical thinking
D Satisfactory	Satisfactory understanding of theory, but with significant shortcomings	Demonstrates limited analytical, technical and writing skills	Limited degree of judgment and independent critical thinking
E Sufficient	Meets the minimum understanding of concepts	Demonstrates sufficient analytical, technical and writing skills	Very limited degree of judgment and independent critical thinking
F Fail	Fail to meet the minimum academic criteria.	No demonstration of analytical, technical and writing skills	Absence of judgment and independent critical thinking

*Adapted from The Norwegian Association of Higher Education Institutions (http://www.uhr.no/utdanning/karakterpanel_1)

The assignment is worth 100 % of the grade of the course.