ITU Computer and Informatics Faculty
BLG 454E Learning From Data, Spring 2018
Term Project

Due 18.05.2018

Prediction of Return in Online Shopping

Suppose that you work as a data analyst in a company. The incurred cost of returns forces your company to take measures to reduce the number of returns without affecting the customer satisfaction. You are asked to solve this problem.

1 Task

For the task, historical data (train set) of one year are known (approximately 481,000 order items) by means of which a model for the prediction of the returns can be learned. The target attribute "returnShipment" of the order item is known here, and it is described by the parameter value "0" in the case "item kept" and the parameter value "1" in the case "item returned". For the purchases of one month (test set) (about 50,000 order items) it is to be assessed in each case whether the item will be returned or not. For this purpose for each order item a prediction is to be made the value of which is within the interval [0,1]. The higher the value, the more probable is the return. The error with respect to the real outcome concerning the return of the order item should be as small as possible.

All the files referred in the term project can be found in ***https://www.kaggle.com/c/blg-454e-term-project-competition/data***

2 Dataset

The historical data contained as well **purchase** and **shipping** data as different **product** and **customer** attributes. The information **return yes/no** was known, too, for the historical data.

For the task anonymized real shop data are provided in the form of structured text files consisting of individual data sets. For the data, in particular the following applies:

1. Each data set is in an individual line that is closed by CR (carriage return, 0xD), or CR and LF (carriage return and line feed, 0xD and 0xA).
2. The first line is structured analog to the data sets but contains the names of the respective columns (data arrays).
3. The header and each data set contain several arrays separated by a semicolon.
4. There is no escape character, and no quota system is used.
5. ASCII is used as character set.
6. There may be missing values. They are coded by the symbol ?.

In concrete terms, only the array names of the attached document **"dataFields.pdf"** in their respective sequence will be used as column headings. The corresponding value ranges are listed there, too.

The training file **"train.txt"** contains all data fields of the document, while the test file **"test.txt"** does not contain the target attribute "returnShipment".

3 Basic Guidelines(Recommendations) of The Project

(a) Data preparation:

- Study carefully the "**verionisleme.pdf**" or "**datacleaning.pdf**" file
- Correcting or cleaning the missing and incorrect data.
- Dimensionality reduction
- Feature engineering: generate/extract new features from data directly improves prediction accuracy/classification performance significantly.

(b) Modeling strategy: Experiment with different methods:

- Binary classification problem (e.g. logistic regression, SVM, Multilayer perceptron, Decision tree, Random Forest).
- Parameter tuning
- You may use any library or built-in functions for the classifiers.

(c) Report the results of your methods using:

- **absolute error** performance measure on training and test set. You can directly get the test error by submitting your results to kaggle system

$$\text{Absolute Error} = \sum_{i=1}^{n} |\text{returnShipment}_i - \text{prediction}_i|$$

Here, returnShipment_i is the information whether order item represents a return (0 means "item kept", 1 means "item returned"), and prediction_i is the predicted return probability for the order item. The team whose error functional reaches the smallest value will win. In case of a tie the decision will be made by drawing lots.

(d) If you want, you can directly implement the referenced paper given in Section 9.

4 Kaggle Class Competition

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

- Term Project Web Site: https://www.kaggle.com/c/blg-454e-term-project-competition/
- Every student has to create a kaggle account and teams have to create their own teams that consists of **at least two** and **at most three people**. Individual teams are not allowed. You must send an email and then we are going to assign a random teammate if you can not find a teammate.
- Team names should be in the following format: *StudentID1_StudentID2_StudentID3*
- Submission format is explained and a sampleSubmission file (**sampleSubmission.txt**) is given in project website.
  - Each sequential number of the order items from the test data must appear here exactly one time.
- The submitted solutions will be evaluated and compared by means of the absolute error on kaggle explained in Section 3 - (c) that is to be minimized.

5 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) and a powerpoint presentation. Total score of your project mark will be calculated as follows:

- Kaggle Result: 50% of your project mark. The public and private leaderboard scores will be averaged.
- Report: 40% of your project mark
- Presentation: 10% of your project mark

Both report and presentation should contain the following sections:

1. (10 points) Introduction (What are you doing and why? Your Kaggle name, team name, score and rank)
2. (35 points) Data sets used (explain the methodology you used for data preparation)
3. (30 points) Methods used (provide exact details of which methods and software you used)
4. (10 points) Results (explain the methodology you used to obtain your results)
5. (10 points) Conclusions
6. (5 points) References

6 Submission Policy

- For the report and code, only electronic submissions through Ninova will be accepted.
- Academic dishonesty, including cheating, plagiarism, and direct copying, is unacceptable.
- Note that **your codes and reports will be checked with the plagiarism tools!**
- Late submissions or those submitted otherwise than according to instructions will not be accepted.

7 Bonus marks

- Top **five** team will be rewarded with bonus marks, respectively, **50pts, 40pts, 30pts, 20pts and 10pts.**, according to the average of the public and private leaderboard scores.
- You may get at most **10pts** from the following criterias:
  - Novel approaches.
  - Clarity of presentation, use of graphics.
  - Depth of understanding.
  - Extended design or analysis.

8 Deductions (-5pts)

- Spelling errors.
- Messiness / lack of proofreading.
- Lack of content.
- Irrelevant / mistaken content.

9 Notes:

1. You are allowed to use only **Matlab** or **Python** programming languages for the implementations.
2. **Reference paper:** I. Bilgen and O. S. Sarac, "***Prediction of return in online shopping***," in Signal Processing and Communications Applications Conference (SIU), 2015 23th, pp. 2577-2580, IEEE, 2015.
3. **Checklist:** "**projectguidelines.pdf**" checklist can guide you through your Term project.
4. It is expected that all team members contribute all the project and report. In the demonstration week, any kind of question that is related to your project can be asked by the instructor!
5. If a question is not clear, please let the teaching assistants know by email **kivrakh@itu.edu.tr** or **cebeci16@itu.edu.tr**.