|  |  |
| --- | --- |
| **P170M109 Computational Intelligence and Decision** | **Date: ……2018 05 30……………………….………..** |
| TOPIC .......6...... | **Students: Matas Valužis ir Deividas Smailys** |

Error based learning

1. Choose the error-based learning model suitable for your dataset  
     
   By inspecting our data set, we have noticed that our target feature RESP is categorical (binary). This led us to select Logistic Regression as error-based learning model, because it handles gradient descent and not having direct breaking point of the decision. Since we are creating our lab works using Python, we chose to use SkLearn Python library.
2. Implement cross validation idea by selecting at least 4 folds.  
     
   We used the KFold method in SkLearn to split the data into 4 folds. Future data will be provided for each of the fold separately unless proven that the quality of data is spread equally in the dataset and splitting is not necessary.
3. Experiment: Select at least two categorical features and at least two continuous features. Perform necessary steps before learning it. Learn the model and write down its formal description (equation). Provide predictions for your target feature using test dataset.  
     
   We have selected two categorical features: PROMOS and STYLES.  
   Learning the model with data as is, gave us predictions for our target feature using test dataset:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Fold Nr. | Mean squared error | Variance score | Accuracy | F1 score |
| 1 | 0.1571 | -0.164 | 84.289 | 0.26 |
| 2 | 0.1609 | -0.101 | 83.905 | 0.30 |
| 3 | 0.1678 | -0.121 | 83.215 | 0.28 |
| 4 | 0.1435 | -0.106 | 85.648 | 0.23 |

Confusion matrix for Fold Number 4:

|  |  |  |
| --- | --- | --- |
|  | 1 | 0 |
| 1 | 112 | 686 |
| 0 | 62 | 4352 |

Similar confusion matrices were generated for other 3 folds. As we can see, there’s a large number of 0 in our target feature data space, therefore we try to normalize data by trimming out the data with 0 in target feature.

Before dropping values, we had a ratio of 1 and 0 of: 0.17 to 0.83;

After dropping values, we have a ratio of 1 and 0 of: 0.4 to 0.6;

Test results with normalized data increased the mean squared error to 0.22 and accuracy dropped to 77.89. In conclusion, data trimming depending on the target feature values will not be implemented further.  
 From the results obtained through raw data model training and predictions we can see that our data quality is spread equally throughout the dataset, therefore further investigation is carried out without using folds.

While using sklearn library chi2 module for stats between features we have calculated p values for the two selected features:  
 PROMOS – resulting in ~0. This might be due to library not handling such low numbers or not working properly. We are assuming that numbers are low enough that they do have a high impact for the model outcome.

STYLES – resulting in ~0. This might be due to library not handling such low numbers or not working properly. We are assuming that numbers are low enough that they do have a high impact for the model outcome.

Formal description:

1. Experiment: Select at least 6 features OR all features you have in training dataset. Perform necessary steps before learning it. Learn the model and write down its formal description (equation). Provide predictions for your target feature using test dataset.  
     
   We have selected six features: 'PROMOS', 'STYLES', 'DAYS', 'REC', 'CC\_CARD', 'WEB'.  
   Learning the model with data as is, gave us predictions for our target feature using test dataset:  
   Mean squared error: 0.141  
   Variance score: -0.090  
   Accuracy: 85.85  
   F1 score: 0.259  
     
   Confusion matrix:

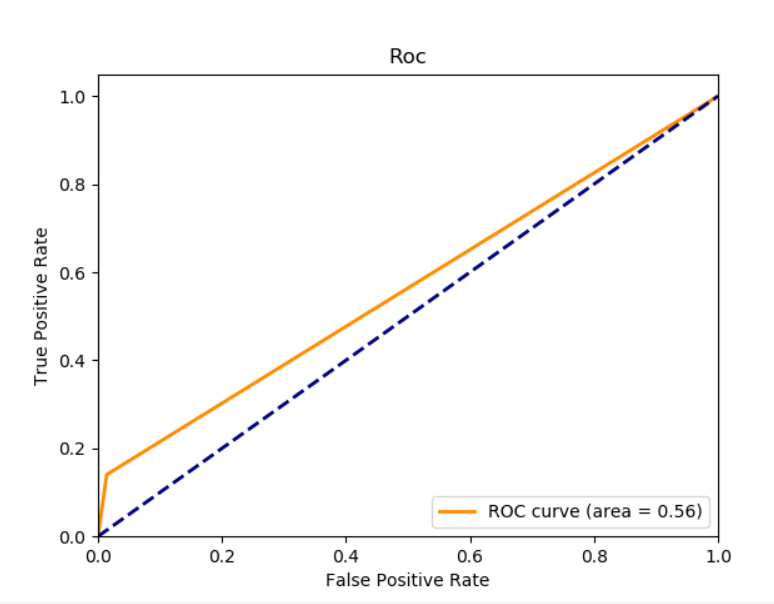
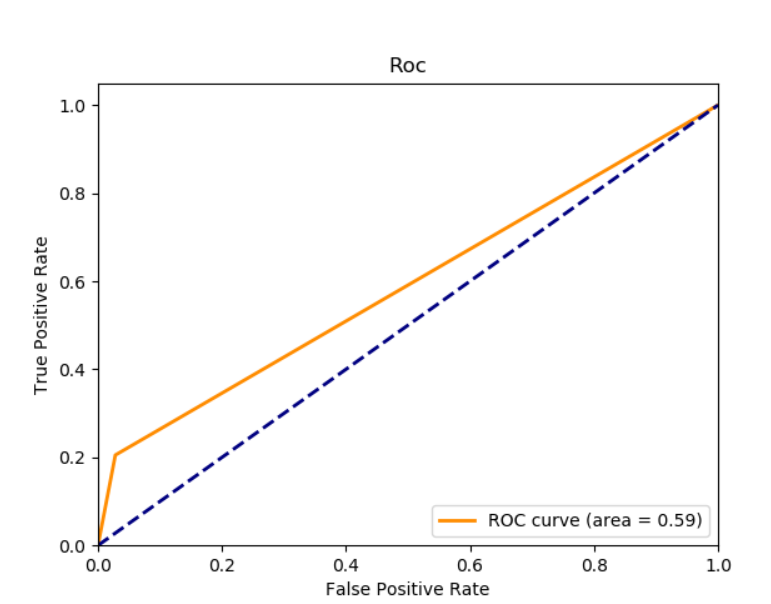
|  |  |  |
| --- | --- | --- |
|  | 1 | 0 |
| 1 | 129 | 669 |
| 0 | 68 | 4346 |

Calculated pValues similarly to the method provided in 3rd section:  
PROMOS: 0   
STYLES: 0  
DAYS: 0

REC: 0

CC\_CARD: 7.71578321e-123

WEB: 1.18951861e-083  
As in the pValues provided in the 3rd section of the report, most of the values are zero or very small. Usually thresholds for p-Values are 1% or 5%. This means that all of the 6 features has impact on the target feature.  
Formal description:

1. For each experiment, find F1 score, provide ROC (AUC) curve and make conclusions about the goodness of model to be learned.   
     
   F1 scores are provided in each section alongside the corresponding results.  
   ROC curve for the experiment with two categorical features:  
     
     
   ROC curve for the experiment with six selected features:  
     
   As we can see, our six-selected feature model performs a bit better, but it is very close to random selection. ROC curve results are not getting aligned with model prediction rate. We might consider reimplementing the Roc curve drawing functionality and investigate possible flaws.
2. Summarize the results to be obtained.  
   Error-based learning models proved to be the most accurate for the given dataset. We managed to obtain 85% accuracy with one of the models that contained 6 selected features from our data set. It’s hard to believe that all the selected features had such a great impact on the prediction outcome due to p-values being very small. Formal definitions also provide some insights about how much a feature can influence the outcome, but it’s not guaranteed, because we cannot see the data distribution and the range for the given coefficient.