MGS 662: ML for IT Managers Project on Blog Data

1)Team Composition and Contribution

Handled the Linear Model and Logit Model for predicting the feedback on given blog data alone. Set R on Mac and installed the required graph package and prediction package

2)Data Set Preprocessing

Experiment 1: Extracted the attributes from column 51 to 60 inclusive which describes the basic features, on the target data feedback in 24 hrs. which is column 281in the given blog data set

Experiment 2: Extracted the attributes from column 63 to 262 inclusive which describes the textual features, on the target data feedback in 24 hrs. which is column 281in the given blog data set

Test data set, February 3rd and 27th with March 14th and 25th dates were selected for prediction

3) Design Issues

Taken the complete train dataset rows. Columns named on the column belonged in original data

Observed on any empty values present, which will affect the observation

Built the models on the following assumptions

Linear Regression: Used R inbuilt linear model, used 1 - (Actual – Expected)/Actual as accuracy calculator, Considered Mean Square Error for calculating deviation

Logistic Regression: Used R inbuilt Logit model, considered the output above zero as one indicating there is feedback on the blog, irrespective of count on the bimodal Logit. Calculated misclassification mean to come to accuracy of model. Considered Mean Square Error for calculating deviation

Built Box plot and linear plots on different input vs target with scatter plot of residuals and expected

4)Empirical Results and Discussions

Case1 Experiment1 Linear Regression Model:

All the features had significant t value except V58 which is a trace back on the number of comments in the time period between T1 and T2

V53,V54 and V56 refers number of feedbacks decreases as the time increases

The percentage of error in the fitted model is 33.19% with the R-square representing roughly 22.54% of the variance found in the target variable feedback count can be explained by the basic feature input variables

The average prediction accuracy after testing on 4 dataset is 32% higher than expected.

Case2 Experiment1 Logistic Regression Model:

The most significant attribute in predicting getting feedback or not is V52 i.e. number of comments in the last 24 hours before the base time feature

Deviance measures fitness of the model. Higher means bad fit. Observed that null deviance shows only the target prediction based on intercept bur residual includes dependent variable. So addition of 8 (52396 – 52388) independent variable decreased the deviance from 68446 to 56467 by 8 degree freedom loss

The average prediction accuracy after testing on 4 dataset is 76.62% of the expected.

Case1 Experiment2 Linear Regression Model:

Generally observed textual features had positive and negative impact on the feedbacks. May be too sentimental words were the once which gets the feedback. For e.g.: V241 had positive impact on feedback while V251 had the negative impact on feedback, both were significant.

The percentage of error in the fitted model is 37.27% with the R-square representing roughly 2.32% of the variance found in the target variable feedback count can be explained by the input textual feature variables

The average prediction accuracy after testing on 4 dataset is 49.6% higher than expected.

Case2 Experiment2 Logistic Regression Model:

The most significant attribute in predicting getting feedback or not is V143 textual feature

Deviance measures fitness of the model. Higher means bad fit. Observed that null deviance shows only the target prediction based on intercept bur residual includes dependent variable. independent variable decreased the deviance from 68446 to 66475 by 187 degree freedom loss

The average prediction accuracy after testing on 4 dataset is 68% of the expected.

Performance: The Model was built quickly for 10 basic features vs 200 textual features. So while considering the columns required we can apply the coherency between the features and select the required attributes

Practical Experience: Graphs display took longer time, R studio halted when tried for building gradient descent function by writing own linear model function. So used the available function

Conclusion: Predicting will the blog get feedback(glm) is more accurate than feedback count(lm)

Link to logs and graphs: https://drive.google.com/file/d/0B1BWPTTXWIpbMS04T0owOWNPOUk/view?usp=sharing

Yashaswini Huskur Lingarajaiah - 50204545