PREDICTING FAKE NEWS OR REAL NEWS

Aina Fatin binti Ahmad Fadzil (16028797)
Nicholas Wong Jing Yan (17027681)

Data Pre-processing

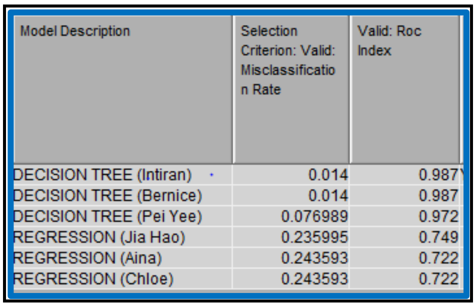
- Followed parsimony concept and build models with increasing complexity
- Insignificant variables to the model such as news_url and tweet_ids are removed
- Missing values are considered to contribute in the prediction, thus no action such as deletion is performed towards those values
- All partitioning was set to 80 for Training and 20 for Validation
- Data was pre-processed to form 8 derived variables, which are domain, ContainQuotation, ContainExclamation, ContainQuestion, NoStopWords, Titlelength, CapsRatio, and TriggerWord

Prediction Models

Regression	Decision Tree
Regression (Aina) **Complexity:** Low **Variables:** T_GivenIDs, T_Retweet, T_Fav, T_AvailDs **Accuracy:** 75.6% **Preparation:** Log transformation	Decision Tree (Intiran) **Complexity:** Low **Variables:** domain, T_GivenIDs, T_Retweet, T_Fav **Accuracy:** 98.6% **Preparation:** -
Regression (Chloe) **Complexity:** Medium **Variables:** T_GivenIDs, T_Retweet, T_Fav, T_AvailDs **Accuracy:** 75.6% **Preparation:** Log transformation, Stratification	Decision Tree (Bernice) **Complexity:** Medium **Variables:** domain, T_GivenIDs, T_Retweet, T_Fav **Accuracy:** 98.6% **Preparation:** Stratification
Regression (Jia Hao) **Complexity:** Complex **Variables:** T_GivenIDs, T_Retweet, T_Fav, T_AvailDs, ContainQuestion, ContainQuotation, CapsRatio, NoStopWords, Titlelength, TriggerWord **Accuracy:** 76.4% **Preparation:** -	Decision Tree (Pei Yee) **Complexity:** Complex **Variables:** domain, T_GivenIDs, T_AvailDs, T_Fav, Titlelength **Accuracy:** 92.3% **Preparation:** -

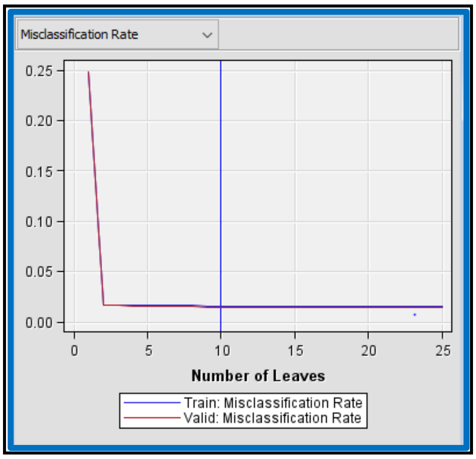
*Log Transformation: To comply to assumptions of normality and rescale variables as parameter estimates sensitive to data sparsity
*Stratification: To prevent underrepresentation as a high proportion of real records compared to fake in hopes to improve accuracy
*According to Shu et al. (2017), headlines are vital in determining fake news, thus linguistic features were extracted from the title

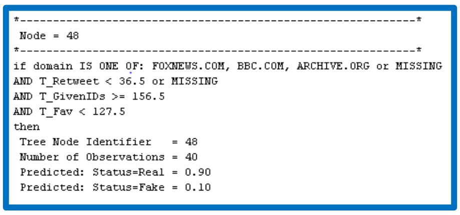
Model Interpretation



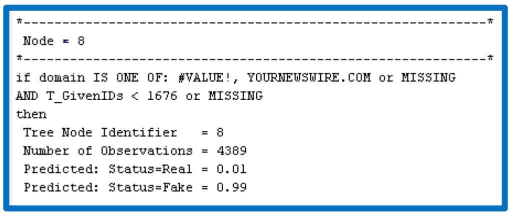
Model Description	Selection Criterion: Valid: Misclassification Rate	Valid: Roc Index
DECISION TREE (Intiran)	0.014	0.987
DECISION TREE (Bernice)	0.014	0.987
DECISION TREE (Pei Yee)	0.076989	0.972
REGRESSION (Jia Hao)	0.235995	0.749
REGRESSION (Aina)	0.243593	0.722
REGRESSION (Chloe)	0.243593	0.722

- Decision Tree (Intiran) has a high ROC, which means that the rate of increase of true positive rate increases faster than false positive rate
- Domain is the most significant variable in the prediction model as it has the highest purity
- The optimal number of leaves that the tree should have is 10 leaves
- Node 48 is the best rule to predict real news. If the conditions of are met, the news is 100% real based on validation
- Node 8 is the best rule to predict fake news. If the conditions of are met, the news is 98.74% fake for validation

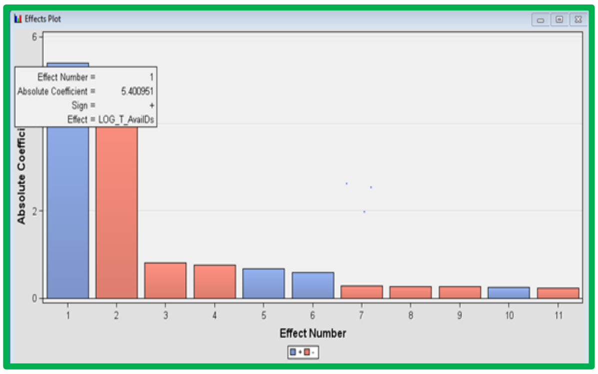




```
*------------------------------------------------------------*
 Node = 48
*------------------------------------------------------------*
if domain IS ONE OF: FOXNEWS.COM, BBC.COM, ARCHIVE.ORG or MISSING
AND T_Retweet < 36.5 or MISSING
AND T_GivenIDs >= 156.5
AND T_Fav < 127.5
then
 Tree Node Identifier   = 48
 Number of Observations = 40
 Predicted: Status=Real = 0.90
 Predicted: Status=Fake = 0.10
```



```
*------------------------------------------------------------*
 Node = 8
*------------------------------------------------------------*
if domain IS ONE OF: #VALUE!, YOURNEWSWIRE.COM or MISSING
AND T_GivenIDs < 1676 or MISSING
then
 Tree Node Identifier   = 8
 Number of Observations = 4389
 Predicted: Status=Real = 0.01
 Predicted: Status=Fake = 0.99
```



- Logistic Regression (Jia Hao) has a high ROC; this means that the rate of increase of true positive rate increases faster than false positive rate
- All the variables included has a p-value lower than 0.05 indicating that they are significant in the prediction of fake news
- T_AvailIDs is the most important variable to the model as it has the highest absolute value

Validity of Best Model

Accuracy = 98.6%
This means that out of all the records in the validation dataset, 98.6% of all observations were predicted correctly.

Precision = 98.66%
This means that out of all the real news (positive) predicted by the model, 98.66% of them were predicted correctly.

Recall = 99.48%
This means that out of all the actual real news, 99.48% of them were predicted correctly by the model.