Online News Popularity

Written by Clément Lajoux

The study

Context:

- Make an IDSS (Intelligent Decision Support System) that can predict if an article will be popular or not (judged popular if the article has more than 1400 shares).

Data used:

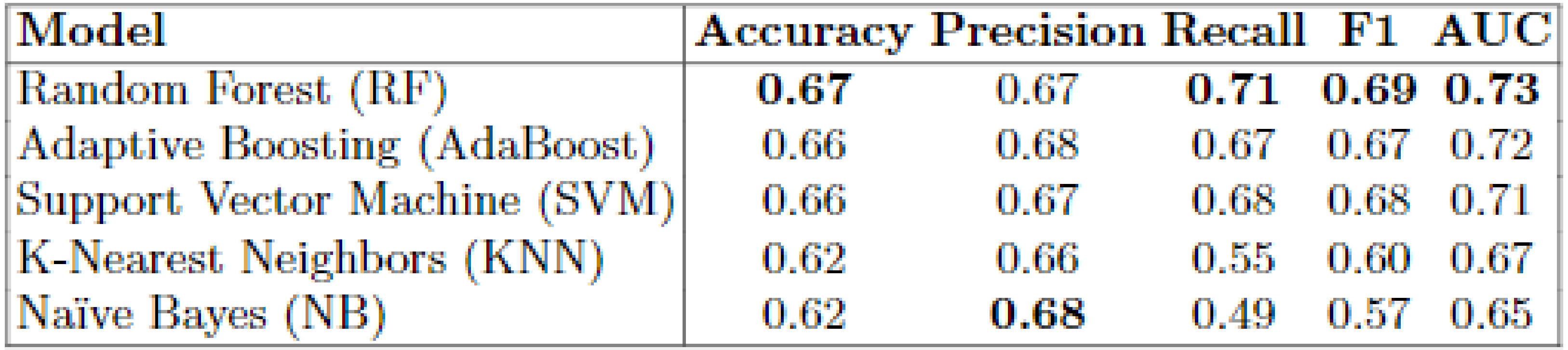
- The IDSS uses 39,000 articles from the Mashable website
- They extract a total of 58 features and 1 target variable from each article.

Analysis of the study

They used the target variable which is the number of shares to create 2 categories, popular and unpopular.

Those two categories are defined by a threshold that can be modified by the user. The threshold used was 1400 shares.

Their best model was a random forest that can achieve an AUC of 73% and an accuracy of 67%



Model	Accuracy	Precision	Recall	F1	AUC
Random Forest (RF)	**0.67**	0.67	**0.71**	**0.69**	**0.73**
Adaptive Boosting (AdaBoost)	0.66	0.68	0.67	0.67	0.72
Support Vector Machine (SVM)	0.66	0.67	0.68	0.68	0.71
K-Nearest Neighbors (KNN)	0.62	0.66	0.55	0.60	0.67
Naïve Bayes (NB)	0.62	**0.68**	0.49	0.57	0.65

Analysis of the study

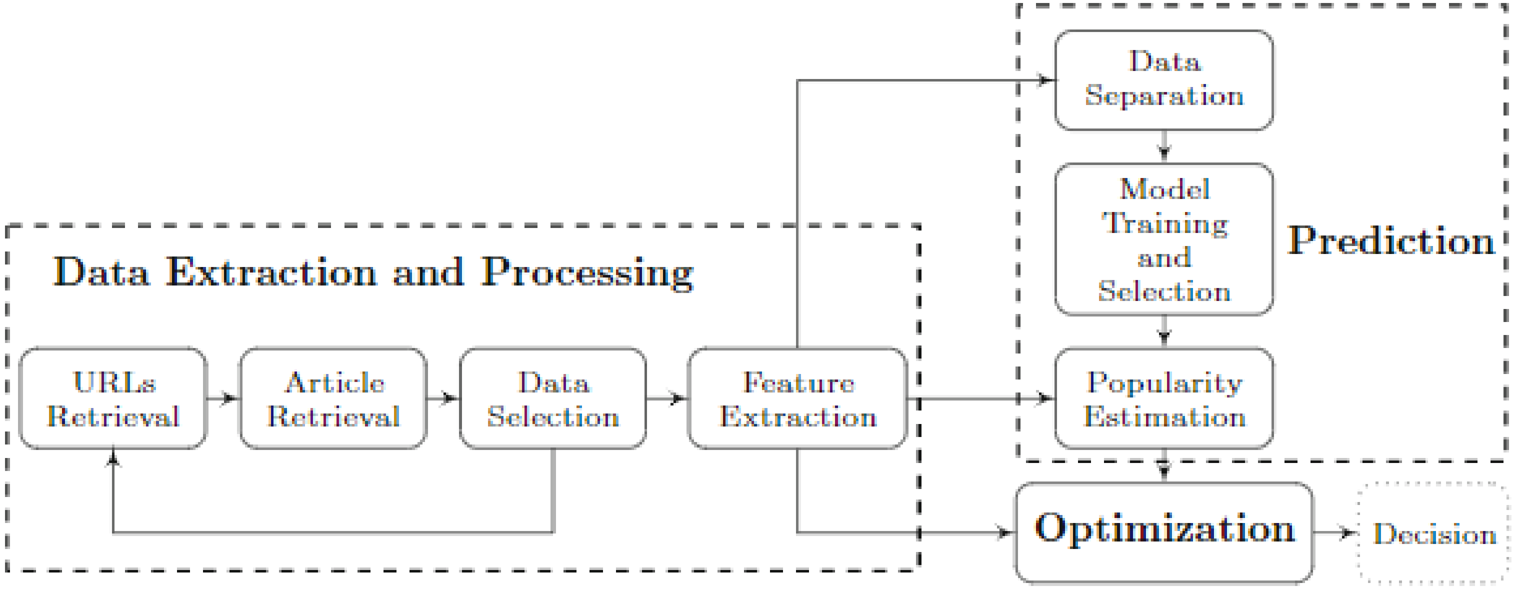
Most of the features come from an NLP (Natural Language Processing) analysis beforehand.

From this analysis they determined features such as the rate positive word, title subjectivity, … They determined 12 features from their NLP analysis.

They determined that the most important features were the keyword-based features followed by Natural Language Processing.

Analysis of the study

The problem I tried to solve was the Prediction part which acts after the Data extraction and Processing as illustrated bellow.



My work based on the data

What have I done?

- A notebook containing the analysis of the study that includes graphs for a better understanding of the data
- A notebook containing the modelization process based on several classification algorithms
- An API that can serve the best model created in the notebook and its client example.

Classification or regression ?

The original target variable was a continuous variable that represented the number of shares.

Since the original study wanted to know if an article will be popular or not, they decided to convert it to a classification problem.

Since I wanted the same type of result to be able to compare it and I think it is more interesting to determine if an article will be popular or not rather than determine the exact number of shares which is impossible.

My results

As far as my analysis goes, I didn’t find any significant correlation between our features and our target variable.

Thanks to a categorization of our target variable, we manage to get a 68% accuracy (higher than the study result) on our modelization.

This accuracy was achieved thanks to a grid search tuning on a gradient boosting classifier.

My hypothesis for the low accuracy is the lack of correlation between all the features and target variables. I think it is still a decent result based on the difficulty of the task.

Credits

Special thanks to:

K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal.

For letting me use their work.