A Comparison of Random Forest and Naïve Bayes Applied to Online Shoppers Purchasing Intention Dataset

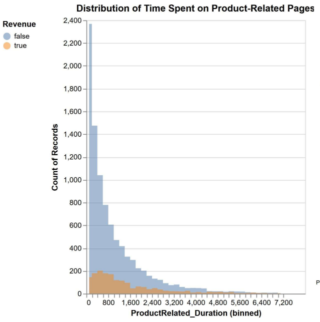
Brenner Swenson

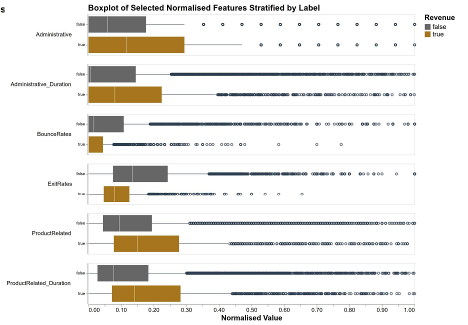
1.0 Description and motivation of the problem

- Marketers spend enormous amounts of time and resources to market and sell products to their consumers. In e-commerce scenarios, consumers reach the checkout point with their items in only to completely abandon the cart 60-75% of the time. [2] As such, being able to predict when a customer will abandon their cart in real-time is of utmost importance to marketers; with this information customers can be provided with offers and promotions that increase the likelihood of a completed transaction.
- Naïve Bayes and Random Forests have shown promise in binary classification in this domain. [1] The results of this project will be compared with those of C. Okan Sakar et al. (2018) who used this same dataset to classify transactions resulting in revenue and those that do not using random forests and **SVM** in addition to more complex methods like **recurrent neural networks**.

2.0 Dataset description and preliminary analysis

- Data obtained from the UCI repository "Online Shoppers Purchasing Intention Dataset", from 2018-08-31.
- The dataset contains 12,330 records that each represent a session on an e-commerce website with 17 features.
- Of the 12,330 sessions, a large majority of them resulted in customers leaving the site (84.5%) without competing a transaction. The other 1,908 sessions ended with a customer purchase.
- For visualisation purposes, in addition to standard normalisation, outliers above the 95th percentile were removed (calculated for each feature independently).
- Nearly all features are highly left-skewed as seen in the boxplot figure (right). This can be observed by the medians in each box marked by the white line.
- Additionally, selected features were split by target value, e.g. 0 or 1, and their distributions visualised. For the selected features there appears to be a relationship between the respective outcomes and the each feature.





3.0 Two machine learning models and their pros and cons

3.1 Naïve Bayes

- Algorithm based on Bayes theorem and is a very prominent probabilistic classification technique used in machine learning [3]
- Aims to select the class that maximizes the **posterior probability** by utilising prior probabilities
- Requires the assumption that each feature is independent from the other; though this assumption can be relaxed in practice without substantial adverse effects [3]

Pros

- The feature independence assumption translates to computational efficiency, making NB attractive for many domains by computing attribute probabilities in parallel. [3] [4]
- When compared to Random Forest, Naïve Bayes provides greater interpretability.
- Uncomplicated to construct and is not sensitive to outliers or low-importance features.

Cons

- Because the independence assumption is so strong, NB-based systems are incapable of using two or more pieces of evidence (features) in conjunction with another. [4]
- Suffers from zero-frequency problem which does not allow NB classifiers to classify instances whose conditional probabilities are zero. [5]

3.2 Random Forest

- An ensemble supervised learning technique based on bagging and random feature selection.
- Uses a number of trained decision trees (base classifiers), taking the majority vote from the ensemble of trees. [6]
- Data is selected randomly for bootstrap samples as it is done in bagging; data is also randomly selected when base classifiers are initialised. [6]

Pros

- Ensembles are often more accurate than any of the single classifiers in the ensemble [6]
- Runs efficiently on large databases and can handle thousands of input variables [6]
- Contains methods for estimating missing data and maintains accuracy when large proportions of data are missing [6] in addition to class imbalance mitigation techniques like **RUSBoost**.

Cons

- The ensemble of trees can become too complex and overfit to the data. A method known as "**pruning**" can be applied to mitigate this and promote better generalisation [7]
- Ensemble outcomes suffer from loss of interpretability compared to a single decision tree
- More computationally expensive to train ensembles than single decision trees.

4.0 Hypothesis Statement

- It is expected that Random Forest with outperform Naïve Bayes for both F1 and AUC as Sakar et al. (2018) produced reported excellent results on the same dataset and did not choose NB as an option.
- This may be due to RF's ability to mitigate class imbalance in the algorithm itself as opposed to NB that requires the dataset to be over/under sampled prior to training.
- Even with **SMOTE** resampling prior to training it is anticipated that RF will still outperform NB when training methodologies like RUSBoost are not used.
- It is not expected that Sakar et al.'s results will be improved upon.

5.0 Training and Evaluation Methodology

- Dataset was split in to 66.66% for training, and 33.33% for testing purposes.
- With class imbalance of (84.5%, 15.5%) it became necessary to use both minority over sampling and random undersampling to bring the ratio of negative to positive examples to 2/1.
- Models were initially trained with MATLAB's hyperparameter optimisation feature to obtain valid values for each hyperparameter. An exhaustive grid search was then applied to each model with each hyperparameter's range including the pre-optimised values.
- K-fold cross validation was used with both models; K was also a parameter included in the grid for each with options ranging from 5 folds to 25 folds.

6.0 Choice of parameters and experimental results

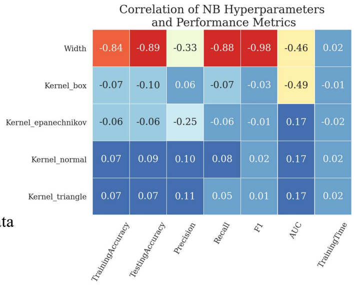
6.1 Naïve Bayes

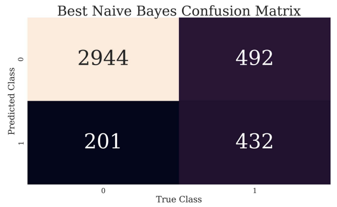
Parameters

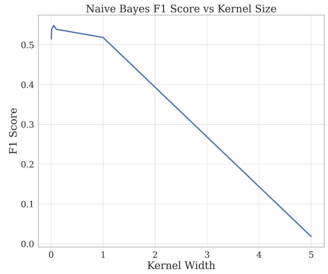
- Triangular **kernel smoothing** was utilised after grid search on all kernel options
- Used a kernel width of 0.05 as well as 20 fold cross validation to achieve best generalisation results

Main Experimental Results

- Various sampling methods were investigated prior to training and running grid search for both models
- Over and under sampling showed the largest effect on **recall**, with a 7% increase compared with imbalanced data
- Various kernel smoothing options were investigated and Epanechnikov, box, and normal options could not outperform the triangular kernel function
- All performance metrics were highly sensitive to **kernel smoothing window width** as evidenced in the correlation matrix to the right







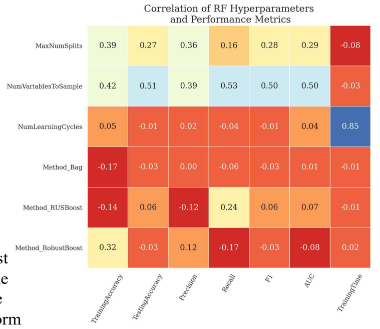
6.2 Random Forest

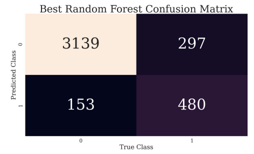
Parameters

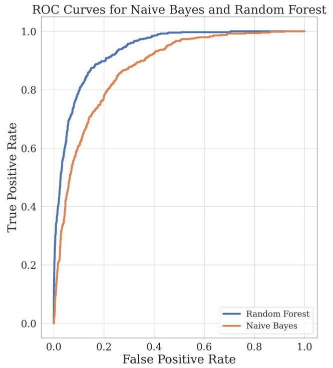
- After a **grid search** of 2,700 parameter combinations, optimal hyperparameters configurations were found that maximise both **F1** score and **AUC**
- Tree depth was limited to 50 splits
- **Minimum leaf size** of 2 was used in conjunction with a minimum parent size of 50

Main Experimental Results

- RF benefitted greatly from the use of the over and under sampled dataset; recall improved by 28%, all else equal, when trained on resampled data.
- **Bootstrap aggregation** (bagging) proved to be the best ensemble aggregation method. Other methods available included those that attempt to mitigate class imbalance issues like RUSBoost. These methods did not outperform resampling prior to beginning the training process.
- **Max. number of splits** was positively correlated with precision but detracted from F1 when increased significantly
- Initially all predictors were able to be chosen for random split, but F1 score showed improvement upon reduction from 75 predictors to 25.
- Optimal number of trees was found to be 50, and resulted in decreased performance as the ensemble grew
- Number of random variables for the initial split remained at the maximum 75 variables; decreasing this parameter resulted in lower performance across all metrics







6.3 Performance Results for Best Models

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1	AUC
Naïve Bayes	85.24%	82.87%	46.53%	67.77%	55.18%	0.857
Random Forest	91.21%	89.21%	62.73%	75.51%	68.53%	0.932

7.0 Analysis and Critical Evaluation of Results

- Congruent with the hypothesis statement, Random Forest outperformed Naïve Bayes for all performance metrics on both over/under sampled data and on unmanipulated data.
- Surprisingly, using oversampled data compared to unsampled training data did not have a material effect on Naive Bayes' performance results. The application of SMOTE did have a small effect on recall but this improvement was offset by a decrease in precision. Conversely, Random Forest was very susceptible to SMOTE with recall improving by 28% on the modified dataset compared to without.
- As generalised Naïve Bayes models tend to produce biased estimate class probabilities, when bias is reduced through training there is an increase in variance [8]. This is likely what is causing the higher error rates, or lower accuracy rates, when compared to the RF model's results.
- With optimised hyperparameters, RF trained 16x faster than the NB model. This is likely due to parallelised sampling that is possible due to the independence of training individual ensemble trees simultaneously [7].
- Contrary to the hypothesis, RF achieved higher test accuracy when compared to literature, but the F1 score still lacks significantly, suggesting the need for further research in to mitigating class imbalance. C. Okan Sakar et al. reported F1 of 0.81 using an oversampled dataset. [1]
- Initial investigations hinted that ensemble size is directly correlated with performance, but after the exhaustive grid search of 2,700 parameter combinations, it is found to be almost completely independent.
- Max number of splits is also very correlated with performance, however this can come at the cost of overfitting. RF is able to mitigate this by the averaging that occurs across all of the random ensemble votes.
- NB's inferior performance may also be a result of highly correlated features that vote twice in the model. [3] This is a direct result of violating the principle of independent features.
- NB classifiers tend too perform poorly with noisy data [9] where majority voting in RF combats this.

8.0 Lessons learned and future work

- It was initially thought that oversampling was required to combat class imbalance primarily for NB due to its effect on prior distributions, but NB benefitted far less from SMOTE than RF did.
- RF F1 score did not compete with literature, so further investigation into sampling techniques is required
- This work did not apply feature selection techniques, where C. Okan Sakar et al. did. Dimensionality reduction, **PCA,** or filter-based feature filtering techniques [1] are of interest in the future.
- Deep neural decision forests have shown promise as a combination of ensemble forest techniques and the benefits of **convolutional neural networks** [7]. Ensemble DNN's are a promising concept worth exploring.

9.0 References

[1]C. O. Sakar, S. O. Polat, M. Katircioglu, and Y. Kastro, 'Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks', Neural Computing & Applications, vol. 31, no. 10, pp. 6893–6908, Oct. 2019, doi: 10.1007/s00521-018-3523-0.

[2]R. K. Rajamma, A. K. Paswan, and M. M. Hossain, 'Why do shoppers abandon shopping cart? Perceived waiting time, risk, and transaction inconvenience', Journal of Product & Brand Management, vol. 18, no. 3, pp. 188–197, Jan. 2009, doi: 10.1108/10610420910957816.

[3]I. Wickramasinghe and H. Kalutarage, 'Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation', Soft Comput, pp. 1–17, Sep. 2020,

[4]D. Xhemali, C. J. Hinde, and R. G. Stone, 'Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages', vol. 4, 2009.

[5]J. Wu, Z. Cai, and X. Zhu, 'Self-adaptive probability estimation for Naive Bayes classification', in The 2013 International Joint Conference on Neural Networks (IJCNN), Aug. 2013, pp. 1–8, [6]V. Y. Kulkarni and P. K. Sinha, 'Pruning of Random Forest classifiers: A survey and future directions', in 2012 International Conference on Data Science Engineering (ICDSE), Jul. 2012, pp. 64–68,

[7]O. Sagi and L. Rokach, 'Ensemble learning: A survey', WIREs Data Mining and Knowledge Discovery, vol. 8, no. 4, p. e1249, Jul. 2018,

[8]LarsenKim, 'Generalized Naive Bayes Classifiers', ACM SIGKDD Explorations Newsletter, Jun. 2005, Accessed: Dec. 10, 2020. [Online].

[9]I. Rish, 'An Empirical Study of the Naïve Bayes Classifier', IJCAI 2001 Work Empir Methods Artif Intell, vol. 3, Jan. 2001.