Sales Prediction Based on e-Commerce Patterns Recognition

@justgrossi
https://github.com/justgrossi/Portfolio.git

***Abstract*—A classifier predicting online purchases was modelled comparing the performances of *Random Forest* and *Logistic Regression*. *Random Forest* was the better performing: Accuracy: 87.6%; Cohen's Kappa: 60.0%; F-Measure: 67.2%. The five most important features when making such predictions resulted to be: page value, traffic type, time spent on an administrative type of page, machine's operating system and pages' exit rates.**

***Index Terms*—Sales Predictions, Random Forest, Logistic Regression, Classification, e-Commerce.**

I. Introduction

Analyzing browsing sessions that culminate in a purchase is crucial within supervised machine learning classification problems due to its direct relevance in predicting consumer behavior. These sessions offer a wealth of sequential data, encompassing various interactions, such as product views, searches, and time spent on pages, leading to a successful transaction. By examining these sequences, supervised learning models can discern patterns and behaviors that distinguish successful purchase journeys from unsuccessful ones. Features derived from such sessions—like specific pages visited, order of interactions, session duration, and referral sources—can serve as valuable predictors. Leveraging this data aids in building predictive models that effectively classify ongoing sessions, allowing eCommerce platforms to identify potential buyers, personalize experiences, optimize conversion rates, and implement targeted marketing strategies based on user behavior. The data set was taken from [1].

II. Related Work

Several studies have explored diverse algorithms in the context of predicting eCommerce purchasing behaviors. Among these algorithms, *Random Forest* and *Gradient Boosting Machine* consistently emerged as high-performing methods showcasing their potential for accurate predictions. Some studies however lacked in-depth discussions on parameter tuning strategies, though, which can potentially have a great impact on classifiers' performance [1]. Notably, *Random Forest* exhibited versatility over standard *Neural Networks* in certain analyses, demonstrating its applicability across a wide range of scenarios [2] [3]. Efforts to address class imbalance in predictive models, such as employing *S.M.O.T.E.* (i.e. Synthetic Minority Oversampling Technique), significantly improved model performance in [4]. In the eCommerce domain, *Random Forest* consistently outperformed other algorithms, proving its efficiency in predicting purchase intentions and guiding targeted marketing strategies [5]. The *Logit* algorithm and *Support Vector Machine* were chosen in [6] to analyse a wine-sales related data set composed of only 1,382 records. Although *Support Vector Machine* exhibited a promising performance, data set's limited size and the lack of insights over the most influential features impacted its effectiveness in providing actionable insights [7]. Also [8] highlighted *Gradient Boosting Machine*'s accuracy in predicting purchasing trends. However, it was *Random Forest* consistently surpassing other methodologies in diverse data sets, showcasing robust performance in different eCommerce contexts [9] [10]. Notably, its strength lies in handling large data sets, categorical features, and on the comprehensible interpretation of variable interactions, thus providing valuable insights into customer behavior [11], [12] [13]. Furthermore, *Random Forest* demonstrated its effectiveness in identifying purchasing patterns and essential attributes for targeted marketing campaigns, aiding in understanding customer behavior, in both [14] and [15]. Overall, *Random Forest* consistently demonstrated superior classification performances, providing accurate predictions and also offering actionable insights into online purchasing behaviors.

III. Methodology

The Cross Industry Standard Process for Data Mining (i.e. *CRISP-DM*) [16] was the adopted methodology. The analysis was facilitated by framework' six stages.

A. *Business understanding*

Predictive models capable of discerning browsing sessions culminating in a purchase offer invaluable insights into consumer behavior and preferences. By accurately identifying the browsing patterns that lead to a purchase, eCommerce platforms gain a deeper understanding of customer needs, enabling tailored marketing strategies and personalized recommendations. Moreover, deciphering the most relevant features influencing purchase decisions empowers businesses to optimize their platforms, enhance user experience, and ultimately drive sales. These predictive models serve as vital tools, allowing eCommerce enterprises to adapt their strategies in real-time, foster customer engagement, and stay competitive in a dynamic online market landscape.

B. *Data understanding*

The data base had 12,330 entries and a total of 18 predictors 4 of which were categorical and 14 numerical. The target label *'Revenue'* identified browsing sessions ending in a purchase. There were no missing values yet the data set was imbalanced with only 15.47% of sessions ending in a purchase (Fig.1).

The data were collected over the span of ten months, and *'operating system'*, *'browser'*, *'region'* and *'traffic type'* appeared to have been anonymised and re-coded with integer values. As for the revenue, the majority of sessions ending in a transaction pertained to returning visitors (Fig.2) and were related to region *'1'* (Fig.3). The peak of the transactions was recorded during the month of November (Fig.4) and the highest number occurred during business days (Fig.5). Both operating system and browser number *'2'* were installed on the machines the highest number of transactions originated from (Fig.6 and 7). Also the predominant type of traffic was *'2'* (Fig.8).

C. Algorithms Selection

Random Forest and *Logistic Regression* were the algorithms of choice. The former was selected as it performed consistently well across all the reviewed related works. *Logistic Regression* on the other hand, wasn't applied as often, therefore, it was tested to verify whether similar performances, as the ones of more sophisticated models, could be achieved.

D. Data preparation

Categorical variables were one-hot encoded to facilitate the application of the minority oversampling technique to resolve class imbalance. As it applies the *K-Nearest Neighbour* algorithm to derive synthetic samples of the minority class, it can only be fed numerical features. Additionally, variables with zero or near-zero variability were removed from the pool of predictors. Finally, test and training set were created adopting a 20/80 ratio.

E. Modelling

In-sample performances were tested with a cross-validation type of procedure on 10 folders. *Random Forest*'s hyperparameters were fine tuned applying a *grid search* strategy. Following are the values that yielded the best in-sample accuracy:

- *'mtry'*: number of randomly selected predictors at each split (Fig.9): 5;
- *'ntree'*: maximum number of trees to grow: 50;
- *'maxnodes'*: maximum number of terminal nodes in each tree: 15.

The out-of-sample performance was then assessed on the test set.

IV. Evaluation

Random Forest scored the highest accuracy of 87.6% (Table I). Despite a lower performance than those recorded in the reviewed literature it was still higher than the no information rate (i.e. *N.I.R.: 84.7%*, *p-value*: 2.52e-16) hence the model was better than a classifier always predicting the majority class. *Cohen's Kappa*, accounting for correct predictions due to chance, was however significantly lower at 60%. With a 56.4% precision the model recorded a fairly low quality of the positive predictions. In addition, not well balanced accuracy, precision and recall meant the model wasn't equally good at predicting both classes. As shown by the confusion matrix in Table II, 242 browsing sessions were false positives and mistakenly classified as ending in a purchase. Class imbalance was only resolved on the training set, therefore, it could be among the causes of such discrepancies compared to the reviewed related works.

TABLE I
Model Evaluation

Model	*Accuracy*	*Sensitivity*	*Specificity*	*Kappa*	*FScore*
RF	87.6%	83.2%	88.4%	60%	67.2%
LR	88%	76%	90.2%	59%	66%

TABLE II
Confusion Matrix

	RF		*LR*	
	References			
Predictions	*No*	*Yes*	*No*	*Yes*
No	1848	63	1886	90
Yes	242	313	204	286

The *R.O.C.* curve was close to the ideal spot on the upper-left corner (Fig.10) and the area under the curve was 85.8%. The classifier identified the following as the top 5 features influencing predictions the most (Fig.11): *'Page Values'*, *'Traffic Type'*, *'Administrative'*, *'Operating System'* and *'Exit Rates'*. *Logistic Regression* recorded similar metrics. The *R.O.C.* curve was fairly close to the ideal spot even in this case (Fig.12), and the area under the curve was 83.1%. The algorithm identified the following as the most important features (Fig.13): *'Page Values'*, *'Month May'*, *'Exit Rates'*, *'Month Mar'* and *'Month Dec'*.

V. Conclusions and Future Work

Predictive analytics benefit online businesses in multiple capacities. A clear understanding of the customer online purchasing behaviour facilitates targeted marketing initiatives. New cohorts could be penetrated and existing ones saturated leading to increased revenue and retention, with customer relationship management and loyalty building strategies being optimized leveraging real-time feedback. *Returning visitors* for instance seem to have higher levels of motivation, and spend more time on website's pages to collect the information they need to inform their decision making. Communication strategies hence could be perfected accordingly, with notifications and/or pop-ups delivering customized recommendations or last-minute deals. *Ad-hoc* promotions and discounts could be implemented as well to combat seasonality. The average order value too could be increased by calibrating pricing strategies and tactics to each customer's behaviour. Different pricing could be displayed to returning customers, and/or when special days are getting closer. By cross-referencing this information with browsing regions, states, addresses, and zip codes. etc. pricing could be geo-located. Considering the relevance of metrics such as *page value* and *exit rate* pages that don’t exhibit acceptable thresholds could be revised. Those with higher average values, on the other hand, expanded or integrated with a stronger ’call to action. More refined purchasing options could be suggested to returning customers who are more motivated and likely to conclude a transaction. Specific actions such as signing up for the emailing list, checking a particular product description more than once, or even adding items to the basket, etc., could function as triggers for automated notifications to propose a targeted discount on that specific item.

References

[1] C. Sakar, Y. Kastro, 2018, ”Online Shoppers Purchasing Intention Dataset”. Available at: https://archive.ics.uci.edu/dataset/468/

[2] R. Kabir, F. Ashraf, R. Ajwad, ”Analysis of Different Predicting Model for Online Shoppers’ Purchase Intention from Empirical Data”, 2019, pp. 1-6, doi: 10.1109/ICCIT48885.2019.9038521.

[3] X. Hu, Y. Yang, L. Chen, S. Zhu, ”Research on a Prediction Model of Online Shopping Behavior Based on Deep Forest Algorithm”, in 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD), 2020, pp. 137-141, doi: 10.1109/ICAIBD49809.2020.9137436.

[4] Z.H. Zhou, J. Feng, ”Deep forest”, in National Science Review, Volume 6, Issue 1, 2019, pp 74–86, doi: 10.1093/nsr/nwy108

[5] K. Baati, M. Mohsil, ”Real-time prediction of online shoppers’ purchasing intention using random forest”, in IFIP International Conference on Artificial Intelligence Applications and Innovations, Springer, 2020, pp. 43-51

[6] P. Bajaj, R. Ray, S. Shedge, S. Vidhate and N. Shardoor, ”Sales Prediction Using Machine Learning Algorithms”, in International Research Journal of Engineering and Technology, 2020

[7] D. Van den Poel, W. Buckinx, ”Predicting online-purchasing behaviour”, in European Journal of Operational Research, Volume 166, Issue 2, 2005, pp 557-575, https://doi.org/10.1016/j.ejor.2004.04.022.

[8] X. Hu, Y. Yang, S. Zhu and L. Chen, ”Research on a Hybrid Prediction Model for Purchase Behavior Based on Logistic Regression and Support Vector Machine”, in 2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD), 2020, pp. 200-204, doi: 10.1109/ICAIBD49809.2020.9137484.

[9] S. Cheriyan, S. Ibrahim, S. Mohanan and S. Treesa, ”Intelligent Sales Prediction Using Machine Learning Techniques”, in 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE), 2018, pp. 53-58, doi: 10.1109/iCCECOME.2018.8659115.

[10] K. Singh, P.M. Booma, U. Eaganathan, ”E-Commerce System for Sale Prediction Using Machine Learning Technique”, in Journal of Physics: Conference Series, Volume 1712, International Conference On Computational Physics in Emerging Technologies (ICCPET), Mangalore, India, 2020

[11] H.S. Seippel, ”Customer purchase prediction through machine learning”, Master’s Thesis, Faculty of Electrical Engineering, University of Twente, Enschede, The Netherlands, 2018, pp. 1–95.

[12] K. Hambarde, G. Silahtaroğlu, S. Khamitkar, P. Bhalchandra, H. Shaikh, G. Kulkarni, P. Tamsekar, P. Samale, ”Data Analytics Implemented over E-commerce Data to Evaluate Performance of Supervised Learning Approaches in Relation to Customer Behavior”, Singapore, Springer, 2019 https://doi.org/10.1007/978-981-15-0035-0_22

[13] H. Pallathadka, E. H. Ramirez-Asis, T. P. Loli-Poma, K. Kaliyaperumal, R. J. M. Ventayen, M. Naved, ”Applications of artificial intelligence in business management, e-commerce and finance”, in Materials Today: Proceedings, 2021, https://doi.org/10.1016/j.matpr.2021.06.419.

[14] A.B. Shaik, S. Srinivasan, ”A Brief Survey on Random Forest Ensembles in Classification Model”, in International Conference on Innovative Computing and Communications, vol 56, Singapore, Springer, 2018, https://doi.org/10.1007/978-981-13-2354-6_27

[15] O. Piskunova, R. Klochko, ”Classification of e-Commerce customers based on data science techniques”, in CEUR Workshop Proc., Vol. 2649, Kyiv, 2020

[16] CRISP-DM, Data Science Project Management. https://www.datascience-pm.com/crisp-dm-2/ (accessed Jun. 09, 2021).

Appendix

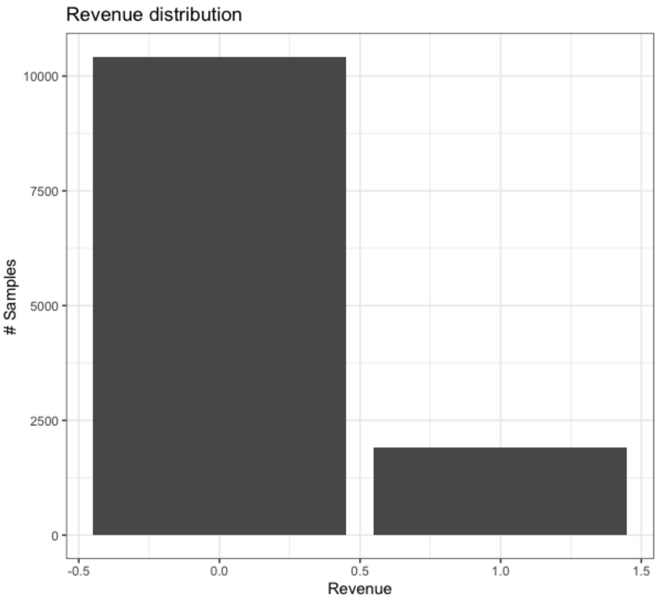

Fig. 1. Revenue distribution

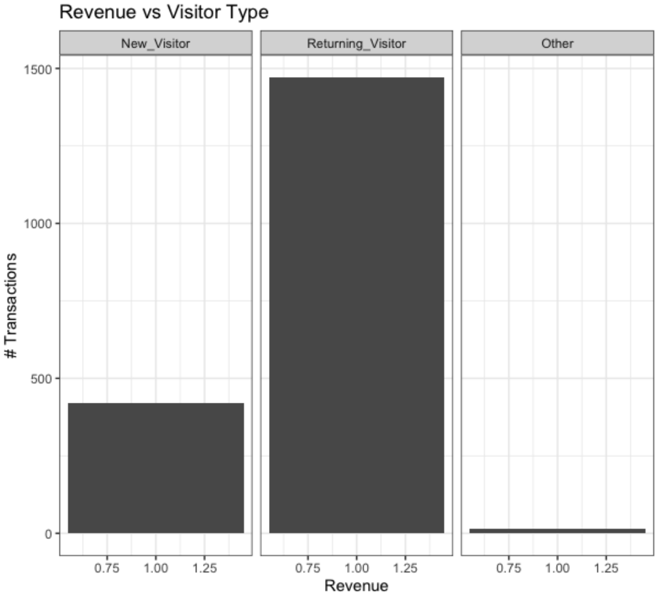

Fig. 2. Revenue vs Visitor Type

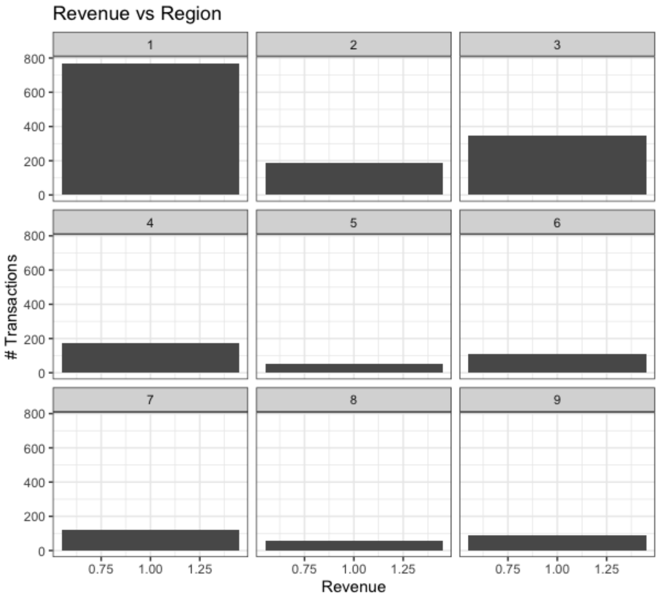

Fig. 3. Revenue vs Region

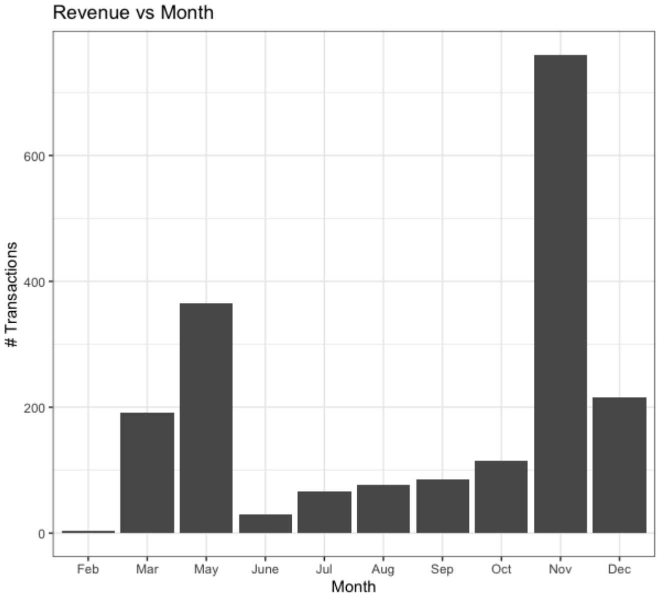

Fig. 4. Revenue vs Month

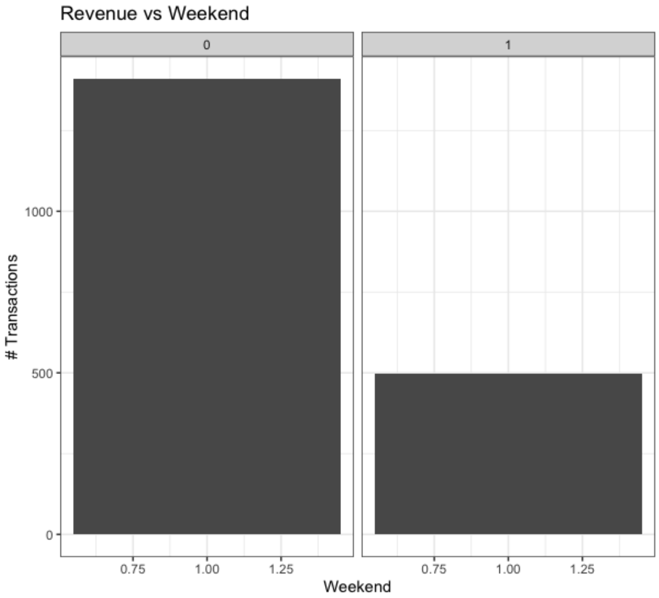

Fig. 5. Revenue vs Weekend

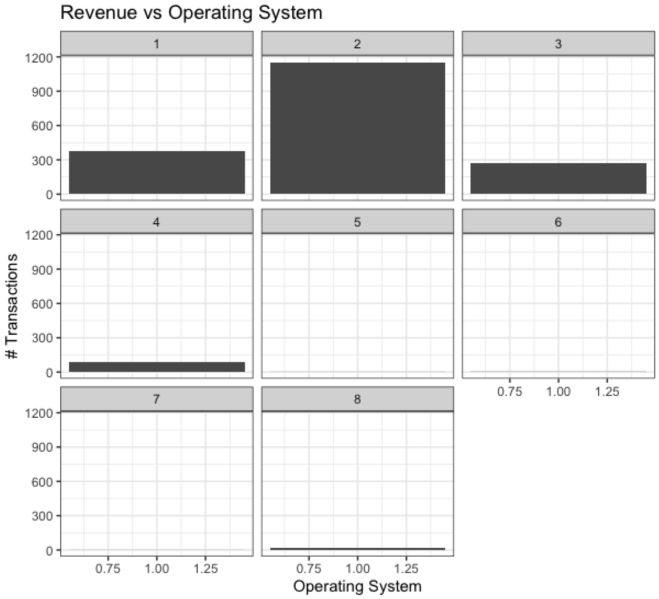

Fig. 6. Revenue vs Operating System

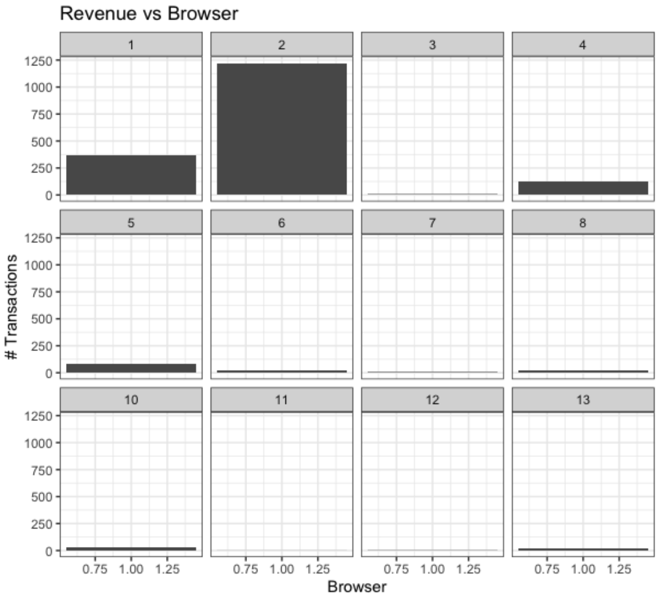

Fig. 7. Revenue vs Browser

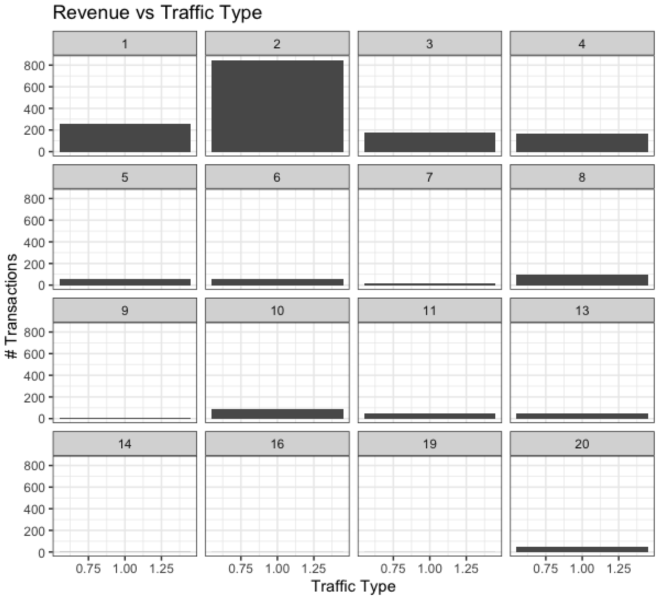

Fig. 8. Revenue vs Traffic Type

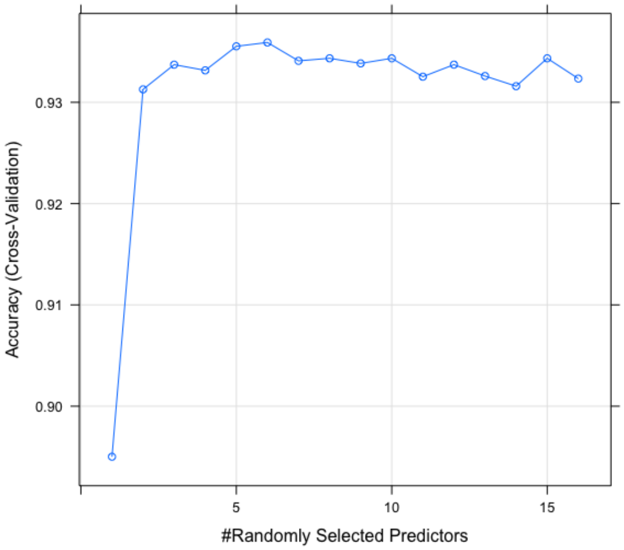

Fig. 9. Random Forest: mtry parameter tuning

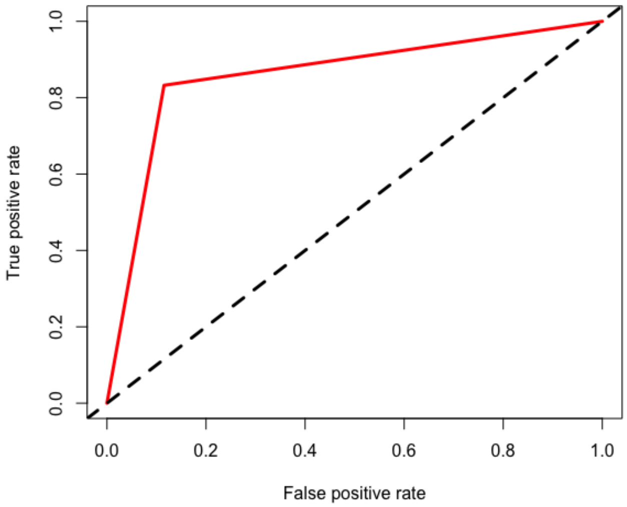

Fig. 10. Random Forest: R.O.C. curve

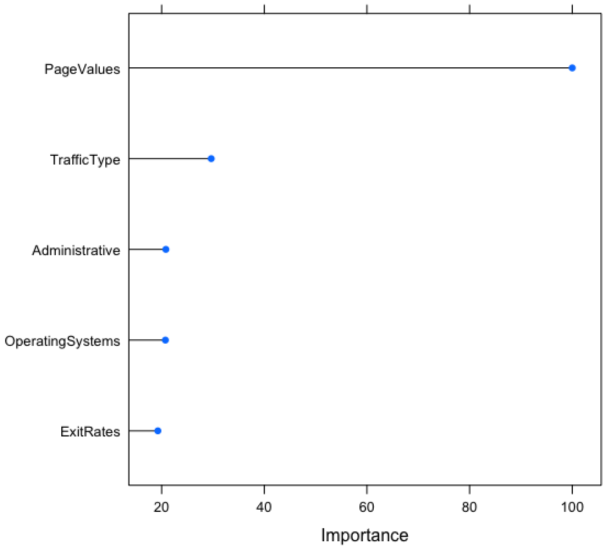

Fig. 11. Random Forest: feature importance

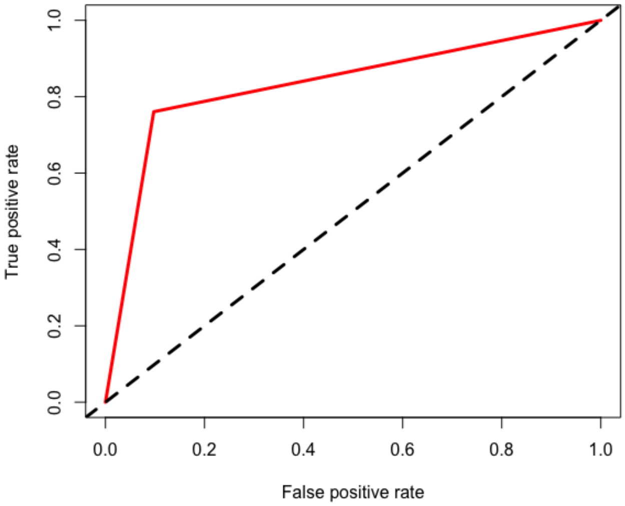

Fig. 12. Logistic Regression: R.O.C. curve

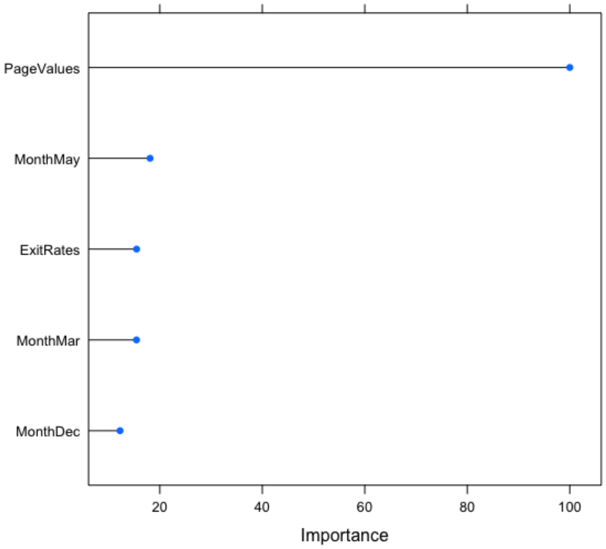

Fig. 13. Logistic Regression: feature importance