Churn Prediction in the Credit Card Industry

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

***Abstract*—A classifier based on *Random Forest* was created to predict churns in the credit card industry (Accuracy: 92%; *Cohen's Kappa: 83.9%; F-Measure: 91.9%*). The three most important features when making those predictions resulted to be: number of transactions; amount of credit bought during the previous 12 months and credit card utilization ratio.**

***Index Terms*—Classification, Random Forest, Decision Trees, credit card.**

I. Introduction

Early churn-detection systems are especially important in subscription-based type of business models to avoid customer and revenue losses. Detecting the early signs of potential defections is paramount in the software as a service industry (*SaaS*). A popular streaming service once dominating the home-entertainment industry for instance made the news after losing a million customers in four months [1]. More traditional competitive landscapes such as the banking industry make no exception. It has long been demonstrated that acquiring new customers is far more expensive than retaining existing ones [2]. Understanding the root cause of defections and being able to anticipate them might constitute thus a competitive advantage. Here is where machine learning and data mining techniques come into play to develop predictive models achieving the highest accuracy possible as to create tremendous impact on the business.

- ***Research Question:*** what is the highest accuracy a classifier predicting credit card churns in this context can achieve?
- ***Research Question:*** which are the most important features in making such predictions?

The data set was retrieved from the *UCI Machine Learning Repository* [3].

II. Related Work

In the financial and investment service industries churns are commonly defined as customers divesting their portfolio below specific thresholds [4], revoking ancillary services or closing their positions altogether [5]. Research in the field however seems to have the tendency to overlook the fundamental issue of class imbalance. Most data mining techniques to properly execute and output reliable predictions require a roughly equal amount of instances for each target label's levels [6]. Less examples of the minority class mean less information the algorithm has to properly learn its characteristics. Only the predominant class is properly mapped and predictions are skewed in its favour as a result [7] [8] [9]. In [10] different churns detecting algorithms were compared but class imbalance was not addressed. *Decision trees* emerged as the best performing methodology even though results cannot be considered reliable. The same approach is adopted in [11] and [12] where authors rely on the potentially misleading accuracy to support their findings in favour of *Decision Trees*. An ensemble type of approach to churns detection is encouraged in [13]. Authors argue that despite being some of the most mature data mining methodologies, it is impossible to identify a single algorithm consistently out performing the others [14].

Another comparison is carried out in [15] with regard to the Chinese banking ecosystem. *Support Vector Machine* emerged as the most accurate technique even though its performance was just above a random classifier with a 50% chance of making a correct prediction. Contrasting results were obtained by [16] also basing their research on the Chinese banking industry and choosing *Support Vector Machine* as the best performing algorithm. Their predictor, after authors carefully tuned its hyper parameters and addressed class imbalance, reached a 99% accuracy. In [17] *Random Forest* was the best in predicting churns with regard to the Brazilian financial industry outperforming *Decision Trees*, *K-Nearest Neighbors*, *Elastic Net*, *Logistic Regression* and even *Support Vector Machine* based models.

TABLE I
Algorithms Comparison - Accuracy

Reference	Technique	Accuracy [%]
[10]	*Decision Tree*	*85.2*
[11]	*K-Nearest Neighbour*	*88.5*
[12]	*Support Vector Machine*	*97*
[13]	*Gradient Boosting*	*79.7*
[15]	*Support Vector Machine*	*60*
[16]	*Support Vector Machine*	*99*
[17]	*Random Forest*	*90*

III. Methodology

The Cross Industry Standard Process for Data Mining (i.e. *CRISP-DM*) [18] was the adopted methodology. Research questions were approached and facilitated by framework' six stages. *R* was the programming language of choice.

A. Business understanding

Financial institutions need predictive models to identify customers at risk of withdrawing from the service. Achieving the highest accuracy possible would support business' long term revenue goals since customer relationship management strategies could be course-corrected when early defection signs are detected. Insights on what are the most important variables signalling potentially departing customers would be invaluable to build customer loyalty and increase their lifetime value.

B. *Data understanding*

The data base was composed of 10,127 entries for a total of 19 independent variables 5 of which were categorical and 14 numerical. The binomial target label *'Churned'* identified credit card cancellations. *'Income level'* was the only positively skewed variable with 35% of customers earning less than 40,000 and 7% more than 120,000 USD (Fig.1). No missing values were present yet the data set was imbalanced with only 16% of customers recorded as churning. Churns were equally distributed across genders and the average client was 46 of age, possessed an entry-level *'Blue'* type of credit card for an average tenure of 36 months. Those possessing a *'Blue'* card also churned in greater numbers (15%), recorded both the average highest credit card utilization ratio (0.29) and number of customer service contacts (Fig.2). They also tended to buy an average higher number of ancillary products (3.85) than any other cohort (Fig.3).

C. *Algorithms Selection*

Decision Trees and *Random Forest* were selected as they consistently recorded strong performances across the reviewed literature and because they work well with mixed data sets (i.e. were both numeric and categorical variables are present). They are also easy to interpret, the decision making process can be visualized, and features can be ranked based on their relative importance. *Random Forest* in particular allows for predictive features to be ranked by means of the *Gini*'s coefficient mean decrease: the higher the score linked to a variable, the higher its importance.

D. *Data preparation*

Class imbalance was resolved applying the synthetic minority oversampling technique (i.e. *SMOTE*) [19]: synthetic duplicates of the minority class were created to counterbalance the majority. Since the technique requires numerical data as inputs, the categorical features were converted with the *'dummycols()'* function of the *'fastDummy'* library. In the resulting balanced data base 48% of records belonged to the minority class and 52% to the majority respectively. An 80/20 split was then performed using R's *Caret* package to create training and test sets. The balance of the target label within the two was preserved.

E. *Modelling*

The tree-based classifier was created and validated with a 10-folder cross validation procedure to test its in-sample performance. Accuracy was used to tune the complexity parameter *'C'* the best value of which corresponds to the maximum achievable in-sample cross validation accuracy. The optimal value resulted to be 0.01 (Fig.4) thus the tree was pruned in correspondence of this value (Fig.5). *Random forest* was cross validated over 10 folders too. This ensemble methodology usually shows increased predictive capabilities leveraging the bagging technique. Numerous trees are created on multiple bootstrap randomly-drawn from the original data base samples. Instances of interest are then classified based on the majority vote resulting from all the trees. Three hyper parameters need tuning:

- ***'mtry'***: maximum number of features randomly considered and used at each split to make a decision. An *'mtry'* value of 10 yielded the highest in-sample accuracy of 97.7% (kappa 95.4%) (Fig.6),
- ***'ntree'***: maximum number of trees to grown within each bootstrap sample. The highest in-sample accuracy of 93.3% (kappa 86.6%) was linked to a value of 200,
- ***'maxnodes'***: the number of terminal nodes for each tree. The highest in-sample accuracy of 93.5% (kappa 87%) was obtain with a value of 29.

IV. Evaluation

Decision Trees and *Random Forest* exceeded reviewed literature's models' performances. Balancing the data set probably contributed in determining better results. When compared against each other, models' performances were almost equivalent as only negligible differences were recorded. *Decision Trees*' accuracy (91.3%) was significantly higher than the no information rate (*NIR* - 51.1%, p-value=2e-16) hence the model was significantly better than a predictor always predicting the majority class. The adjusted version of the accuracy *Kappa*, accounting for correct predictions due to chance, resulted to be quite high as well (82%). With a 92.2% precision the model recorded a fairly high quality of the positive predictions. Similar and well balanced accuracy, precision, and recall values meant an healthy confusion matrix hence a model equally good at predicting both classes. As a matter of fact the *F-Score* was 91.4%. The *R.O.C.* curve closely resembled the ideal spot on the upper-left corner (Fig.7) and the *A.U.C.* was 91.3%. *Decision Trees* ranked as the top 5 features (Fig.8): amount of credit bought during the previous 12 months, credit card utilization ratio, credit card limit, credit card type, and gender. *Random Forest*'s accuracy was slightly higher (92%) and much better than a model always predicting the majority class (*N.I.R.* 51.1%, p-value=2.2e-16). With a 91% precision, even in this case the confusion matrix showed the model was equally good at predicting both classes. The *F-Score* was 91.9%. The *R.O.C.* curve was extremely close to the ideal spot (Fig.9) and the *A.U.C.* was 92%. *Random Forest* considered the top 5 features to be (Fig.10): total number of transactions and total amount of those transactions, revolving balance on the credit card, Variation of the transacted amount Q4 over Q1 and Variation of the number of transactions Q4 over Q1.

TABLE II
Confusion Matrix

	Decision Tree		*Random Forest*	
	References			
Predictions	*No*	*Yes*	*No*	*Yes*
No	1540	130	1552	119
Yes	160	1497	148	1508

TABLE III
MODEL EVALUATION

Model	*Accuracy*	*Sensitivity*	*Specificity*	*Kappa*	*FScore*
DT	91.3%	90%	92%	82%	91.4%
RF	92%	92.7%	91.3%	83.9%	91.9%

V. CONCLUSIONS AND FUTURE WORK

A classifier based on *Random Forest* was built to predict customers at risk of churning the credit card service. The recorded accuracy was 92% with a Cohen's Kappa of 83.9% and an F-Measure of 91.9%. Total number of transactions during the previous 12 months and their amount; revolving balance on the credit card; variation of both the transacted amount and of the number of transactions Q4 over Q1 emerged as the most important features when making predictions. Future research should consider expanding the data sets including both more records and features. Class imbalance should be considered as well, and balancing strategies other than the minority oversampling techniques could be explored, too.

REFERENCES

[1] N. Sherman, J. Claiton, Netflix loses almost a million subscribers, 2022, BBC News, Available at: $https://www.bbc.com/news/business-62226912$

[2] J. Sterne, 2002, Web metrics: proven methods for measuring web site success, John Wiley and Sons, New York, p. 283.

[3] UCI Machine Learning Repository. Available at: https://archive.ics.uci.edu/ml/datasets.php

[4] J.F. Adolfsen, F. Kuik, E. M. Lis, T. Schuler, The impact of the war in Ukraine on euro area energy markets, European Central Bank Economic Bulletin, Issue 4/2022. Available at: www.ecb.europa.eu/pub/economic-bulletin/focus/2022/html/ecb.ebbox202204$_0$1 68ef3c3dc6.en.html

[5] N. Glady, B. Baesens, and C. Croux, Modeling churn using customer lifetime value, Eur. J. Oper. Res., vol. 197, no. 1, pp. 402–411, 2009, doi: 10.1016/j.ejor.2008.06.027.

[6] D. Van den Poel and B. Larivière, Customer attrition analysis for financial services using proportional hazard models, Eur. J. Oper. Res., vol. 157, no. 1, pp. 196–217, 2004, doi: 10.1016/S0377-2217(03)00069-9.

[7] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, On the Class Imbalance Problem, 2008, Fourth International Conference on Natural Computation, vol. 4, pp. 192–201. doi: 10.1109/ICNC.2008.871.

[8] P. Xenopoulos, Introducing DeepBalance: Random deep belief network ensembles to address class imbalance, 2017, IEEE International Conference on Big Data (Big Data), pp. 3684–3689. doi: 10.1109/BigData.2017.8258364.

[9] H. He and E. A. Garcia, Learning from Imbalanced Data, IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[10] I. Kaur and J. Kaur, Customer Churn Analysis and Prediction in Banking Industry using Machine Learning, 2020, Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), pp. 434–437. doi: 10.1109/PDGC50313.2020.9315761.

[11] A. Bhatnagar and S. Srivastava, A Robust Model for Churn Prediction using Supervised Machine Learning, 2019, IEEE 9th International Conference on Advanced Computing (IACC), pp. 45–49. doi: 10.1109/IACC48062.2019.8971494.

[12] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. Ch. Chatzisavvas, A comparison of machine learning techniques for customer churn prediction, 2015, Simul. Model. Pract. Theory, vol. 55, pp. 1–9, doi: 10.1016/j.simpat.2015.03.003.

[13] H. T. N. Thia, V. N. Thao, Building a proper churn prediction model for Vietnam's mobile banking service, 2022, Institute of Advanced Science Extension (IASE), doi: 10.21833/ijaas.2022.07.014

[14] G. Wang, L. Liu, Y. Peng, G. Nie, G. Kou, and Y. Shi, Predicting Credit Card Holder Churn in Banks of China Using Data Mining and MCDM, 2010, IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 3, pp. 215–218. doi: 10.1109/WI-IAT.2010.237.

[15] J. Zhao and X.-H. Dang, Bank Customer Churn Prediction Based on Support Vector Machine: Taking a Commercial Bank's VIP Customer Churn as the Example, 2008, 4th International Conference on Wireless Communications, Networking and Mobile Computing, Oct. 2008, pp. 1–4. doi: 10.1109/WiCom.2008.2509.

[16] Benlan He, Yong Shi, Qian Wan, Xi Zhao, Prediction of Customer Attrition of Commercial Banks based on SVM Model, P*R.O.C.*edia Computer Science, Volume 31, 2014, Pages 423-430, ISSN 1877-0509, https://doi.org/10.1016/j.p*R.O.C.*s.2014.05.286.

[17] R. A. de Lima Lemos, T. C. Silva, B. M. Tabak, Propension to customer churn in a financial institution: a machine learning approach. Neural Comput and Applic 34, 11751–11768 (2022). doi.org/10.1007/s00521-022-07067-x

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

[19] N.V. Chawla, K.W. Bowyer, L. O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, 2002, Journal of Artificial Intelligence, Research Volume, pp 321–357.

APPENDIX

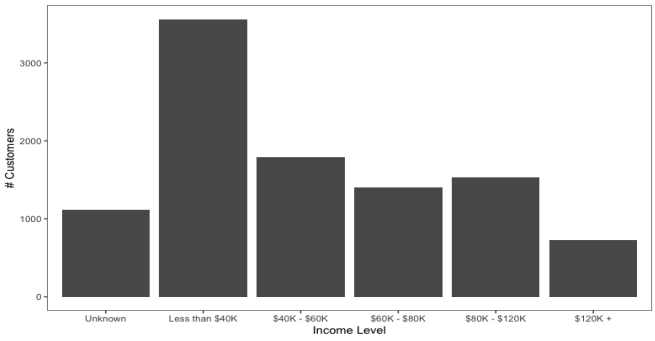

Fig. 1. Income Level Distribution

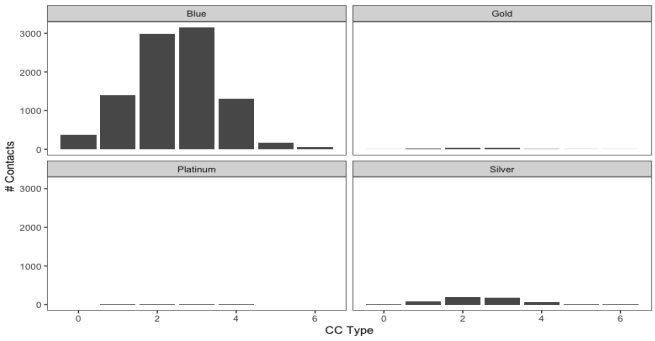

Fig. 2. Number of Contacts vs CC Type

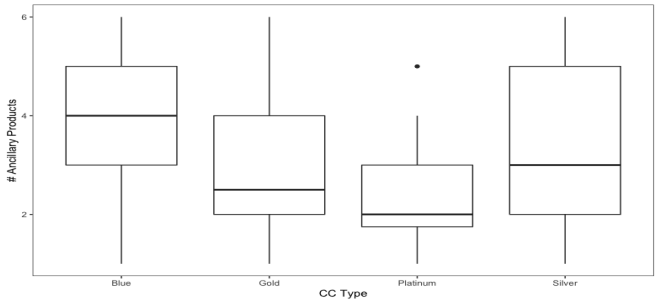

Fig. 3. Ancillary Products vs CC Type

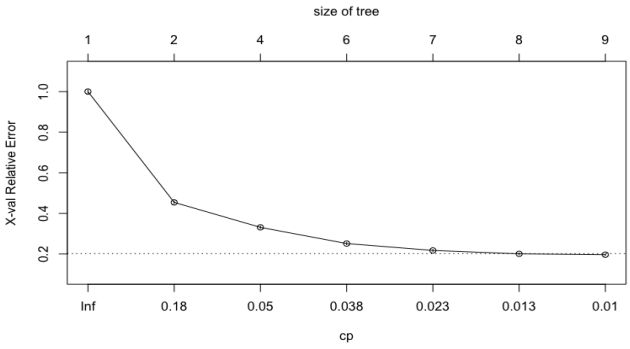

Fig. 4. Decision Tree - Complexity Parameter *'C'* In-sample accuracy

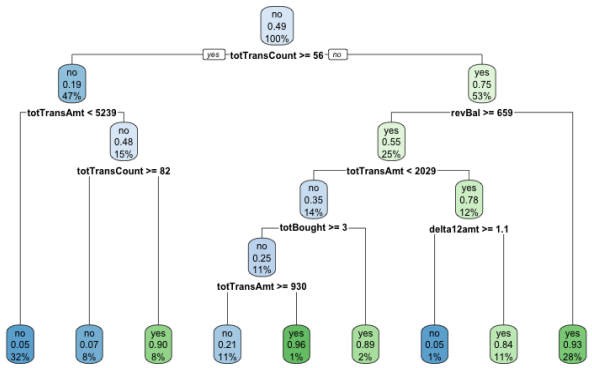

Fig. 5. Decision Tree - Pruned Tree

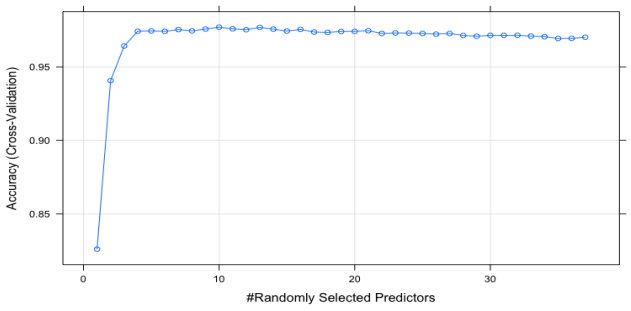

Fig. 6. Random Forest - *mtry* parameter

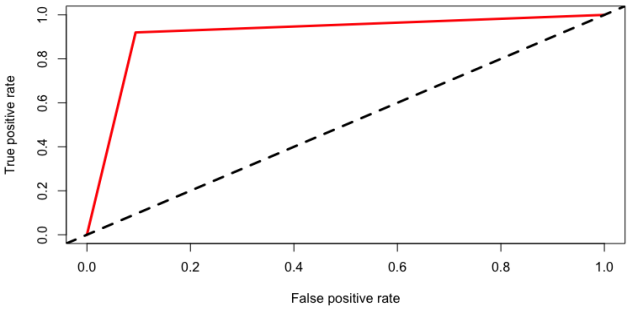

Fig. 7. Decision Tree - *R.O.C.* Curve

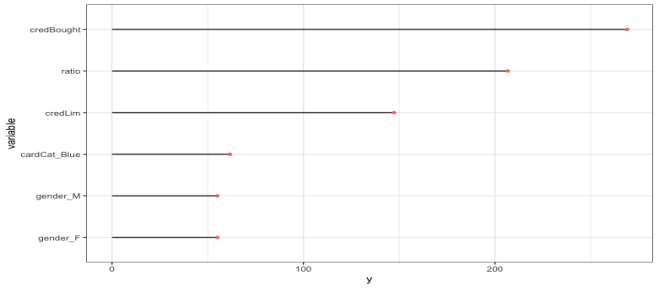

Fig. 8. Decision Tree - Features Importance

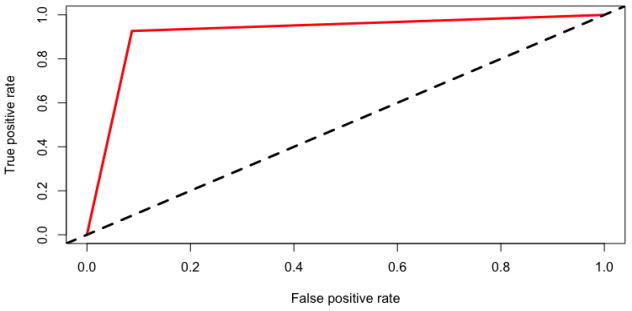

Fig. 9. Random Forest - *R.O.C.* Curve

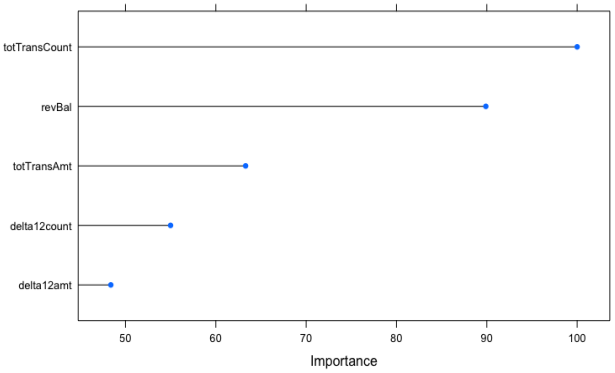

Fig. 10. Random Forest - Features Importance