Artificial Intelligence Research and Development
M. Villaret et al. (Eds.)

doi:10.3233/FAIA210129

A Customer Churn Detection Model for the Pay-TV Sector

Vicente LÓPEZ [a], Rebeca EGEA [b] Lledó MUSEROS [a] and Ismael SANZ [a]
[a] *Universidad Jaume I, Spain*
[b] *Mirada TV, Spain*

Abstract. The business environment today is characterized by high competition and saturated markets. Pay-tv platforms there are not an exception. Because of that, the cost to acquire new customers is much higher than the cost of retaining the existing customers. Therefore, it is important for Pay-TV platforms to keep controlled the Customer Churn. Therefore, the paper studies existing models used to predict Customer Churn in other context -like telecommunication companies customer Churn-and adapts them to the Pay-TV context. Another big problem faced in the paper is the fact that, in the data set udes in the paper there are not personal metrics, which are indispensables to solve the problem. Therefore this approach has defined new metrics in order to be able to predict customer churn.

Keywords. Data mining, Imbalance Classification Problem, Customer churn prediction, Pay-TV platforms.

1. Introduction

We live in a competitive world in which most of services companies have to face the problem of customer churn. Pay-TV platforms are not an exception. The competitivity in this field has grown up, and now it is more expensive to get new customers. Losing customers always means a loss of revenue/profit to the company, but if we consider also the growing costs of getting new customers, the loss can be unaffordable for the company and this can lead the company to the bankrupt.

In Pay-TV paradigm, we can adopt the same definition that was given in [1]: ”Churn is defined to be the activity of customers leaving the company and discarding the services offered by it due to dissatisfaction of the services and/or due to better offering from other providers”. In this approach, our goal is to detect customers with high risk of churn in order to be able to take the necessary actions to prevent it.

There are studies about churn in the field of the telecommunication companies [2], e-commerce [3] or even general studies to solve the problem in general [4]. Different algorithms have been studied to build a good model to solve the problem, and in general, decision trees models have showed better results than other models, specially those used with boosting. Also, there are another good models that could be useful in our context like neural networks or linear regressions [5].

Solving the churn problem has to manage the class imbalance. A big variety of solutions have been proven as useful for solving this problem in some contexts [6]. In

this case, different methods has been tested (undersampling, oversampling and one-class SVM classifiers).

Also, a big difference in the process of predicting customer churn in the Pay-TV context with respect to previous works is that the dataset does not contain any personal metric. The method presented in this paper faces all these problems by developing a decision tree with boosting and using specific Pay-TV metrics as time spent viewing Netflix, number of dispositives used by the user, tate of content viewed -considering the series and movies of the topics that the user has ever seen-, number of subscriptions levels changes and rate of content viewed entirely. The paper demonstrates that the model shows equal or better results that a combination of models using Stacking. The rest of the paper is structured as follows, first a review of works related with churn prediction in other areas are presented, then in section 3 the methodology carried out in this paper is explained and then the model defined in presented. Section 5 analysis the results and then conclusions and future works are outlined.

2. State of Art

In [7] a study about the elaboration of a model capable of predicting Customer Churn inside the telecommunication field is presented. In this study, 4 metrics groups were defined: Customer Demography -personal metrics of the customer-, Bill and Payment -payment behavior-, Call Detail Record -customer behaviour in the company services- and Customer Care Service -customer satisfaction with the company-. This model has inspired the model presented in this paper, but in the Pay-TV there is no data about the group Call Detail Record-. Therefore, instead of these information the model presented in this paper uses another type of information, called View Detail Record which includes those metrics that define the user behavior within the platform. The new category represents the same idea as the group Call Detail Record of [7], thus respecting the chosen metric structure. [8] presents a study about Customer Churn in mobile market, which uses 5 metrics groups: Demographics, Cost, Features/Marketing, Usage Level and Customer Services. By grouping the categories Cost and and Features/Marketing in a set, the result is a set of metrics very similar to the ones used in this paper.

Many studies have been done about the algorithms that can be use for predicting Customer Churn [9,10,11]. [12] presents a general summary about algorithms performance in Customer Churn prediction, and the results show that the algorithms with higher performance are Neural Networks, Decision Tree and Linear Regression. [7] predicts Customer Churn in the telecom paradigm, and it demonstrates that Decision Tree model always surpasses the Neural Network model in the prediction of churn.

Every company, to be able to perdure, needs the number of customers greater than the customers churned since otherwise the company would lose profits very quickly and would end up in bankrupt. It is because of that the Customer Churn is lower in relation with the total number of customer along the company life and that makes our dataset very imbalanced. Work with an imbalance set always causes problems [13]. Trying to solve imbalancement can cause overfitting, making the model accuracy decreases dramatically along with their capability of generalization. Class Imbalance is a very present problem in Customer Churn prediction, and many of the known techniques for solving it have been explored [5]. Among the methods proved useful it is possible to mention

oversampling, undersampling and boosting, which have shown a clear improvement in the model accuracy. These are the methods tested in the model presented.

3. Methodology

The Knowledge Discovery in Databases, also known as KDD, is defined as the ”non trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns of in data”. The problem faced in this work is to identify each customer as ”potentially churner” or ”potentially non churner” at the moment where the model is executed. Then, the KDD function for our problem is defined as a classification problem. Having the correct data is as important as having the correct method [14], so the first step is the adquisition and preparation of data.

3.1. Data Acquisition

In this work an actual dataset, with 300.000 customers entries along two years is used, where around 80% of the customers are non-churned and the rest are churned. The dataset belongs to a private Pay-TV company, called Mirada TV which is a leading provider of cutting-edge digital TV technology, committed to future-proofing the platforms of operators and broadcasters worldwide [1]. Experts from the company have collaborated actively in this model. Taking into account that the dataset comes from real clients, it is important to mention that very specific details of the dataset are not going to be revealed. Nevertheless, it is possible to define the groups of information used in this work, which are the following:

- **Device information:** Information about the hardware that the client is using to access the services of Mirada TV. This group of metrics can determine the economical level of the customer.
- **Bill and Payments:** Purchases and other transactions than the customer does inside the application. It can determine the satisfaction of the customer with the services, along to its economical level.
- **TV Detail Service:** How the customer uses the platform and how much he uses it. It can determine the level of satisfaction with the products of Mirada TV and the level of consumption that the customer has.
- **Errors:** Errors ocurred in any session of a customer. Errors can affect directly the customer’s view of the product. According to the Mirada TV marketing commercials, the errors may be directly related to customer churn.

Notice that, even having the fact that there are sensible information about the clients, there is not information describing the client (i. e. the age, gender of the customer, his/her economical status or laboral situation, offers from competitors, etc). This information is very significant to correctly solve this problem [7], and it represents an additional problem to deal with.

It is also important to point out that there are two different types of metrics in the dataset, which are:

[1] https://www.mirada.tv/about/

- **Variables which are dependent of the time:** There are some metrics which are dependent of the time (i.e the errors per day). Therefore in the model it is important to define a temporal window and these type of metrics would change depending on the temporal windows and their size defined.
- **Variables which are independent of time:** These metrics are totally independent of time (i.e the day the client signed up for the application) and therefore they do not depend on the temporal window defined.

3.2. Exploratory Data Analysis and Data Preparation

One of the problems that it is necessary to address in this model is caused by the very low ratio of clients that leave the platform. This problem generates a very high class imbalance problem. Section 4.2 explains this problem has been managed in this model.

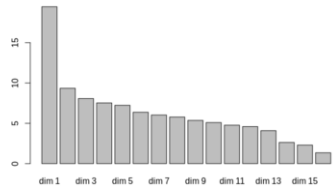


Figure 1. Variation explained for each dimension of the PCA

As explained in the previous section, due to the uses of some time dependent metrics it is necessary to define a temporal window to calculate them. The temporal window used has been defined of 6 months. Also, the dataset is reduced by eliminating the metrics that have a correlation higher than 0.9 to another metric by doing a Principal Component Analysis (PCA) (without PCA the computational cost of the model was too high because the big number of features). The result of the PCA is shown in 1. The results show that, according to the elbow method, there are two reasonable options: keep only one dimension (conserving only 20% of the variance) or keep 5 dimensions (thus conserving 51% of the variance).

4. Model

As presented in the litterature section, there are several models and techniques that have been proved useful to predict customer churn in other areas. Therefore, different alternatives -including the algorithms used and the way to process the data- have been tested to solve the problem in the area of the Pay-TV platforms. It is important to clarify that the model has been implemented using the library scikit for Python [2].

[2]https://scikit-learn.org

4.1. Techniques

In order to build the model, one classification algorithm has to be selected. In [1,7,8,12] different techniques were proved useful in similar problems. To select the technique with which to build the model, several algorithms were tested using a smaller set of the dataset, and the algorithm which yields better results has been selected.

Specificaclly next techniques were tested: neural networks (NN), K-neighbors with the variants of centroids (KNCn) and with principal componen analysis (KNCa), support vector classification (SVC) and also SVC with its linear variant (SVCL) and Nu-support vector classification (NuSVC), one class predictor (OCP), decission trees (DT) and DT with boosting based in gradients (GBC) and histograms (HGBC), and finally logistic regression (LR), and LR with crossvalidation (LRCV). All of them were tested with the defaults parameters sets by scikit-learn [15]. More details can be found in [16].

The dataset used in these experiments was a balanced set with 10.000 churned customers and another 10.000 no churned customers, selected randomly from the original dataset, therefore the dataset has been undersampled. This new dataset is only used in the scope of this section. The same partition was used for all methods, with 70% of data for training and 30% for test. Each test was executed 10 times, and the mean of the scores are shown in figure 2. The results shows that most of the algorithms exceed the 75% of score. As HGBC was the method with better results, this was the one selected to build the model.

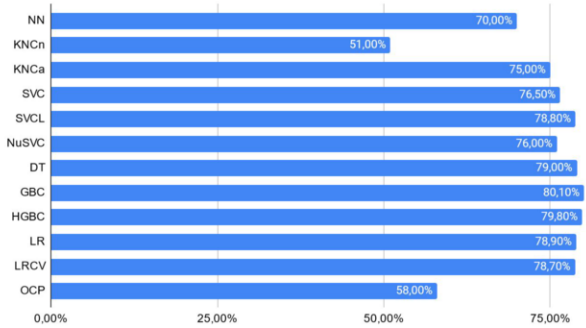


Figure 2. Results of initial test methods.

4.2. Class Imbalance Problem

As mentioned before in this paper, our original dataset is very imbalanced because the number of clients abandoning a platform is very low in comparison with the ones that remains as clients (around 80% are non-churn clients). Therefore the original dataset has a predominant class and the model will learn to predict very well this class but not the other, and for the companies the ”churn class” which is not the dominant is the most important one. There are several techniques that can apply apllied to solve this problem [13]. In this work two different aproximations has been tested: oversampling and undersampling. Both techniques have been applied to the original dataset using the HGBC algorithm for learning. In both cases the experiment was done 10 times. In average, over-

sampling obtains 78% score and undersampling obtain 77% score. Therefore oversampling was selected as the technique to be used in the final model of this work.

4.3. Oblivion Modeling

We humans do not remember the things that happened today with the same intensity as those that happened a week ago. Therefore, metrics that are time dependent are susceptible to be forgetable. Therefore, the model presented in this paper has tried to model the fact that people forget things by defining a "oblivion model". The transformation f proposed is an exponential cubic function that, given a day t_i and the value of that metric in that day $m_{j,i}$, the new value is calculated as follow

$$f(m_{j,i}) = m_{j,i} \cdot e^{\left(\frac{t_i - t_{max}}{d}\right)^3} \tag{1}$$

where d is the half of the days that the interval to take into account has and t_{max} is the value of the max day (normally, the integer value of today). By fixing the t_{max} as 1000, d as 30 and $m_{j,i} = 1$ for all i and for a given j, the resulting function showed in figure 3 is got.

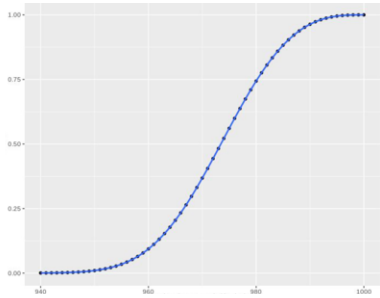


Figure 3. Function of the oblivion model.

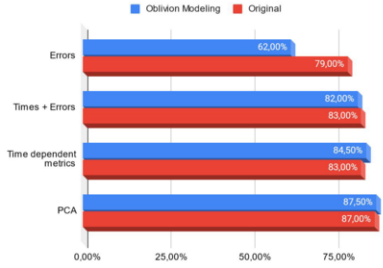


Figure 4. Comparation between the original model and with the oblivion model.

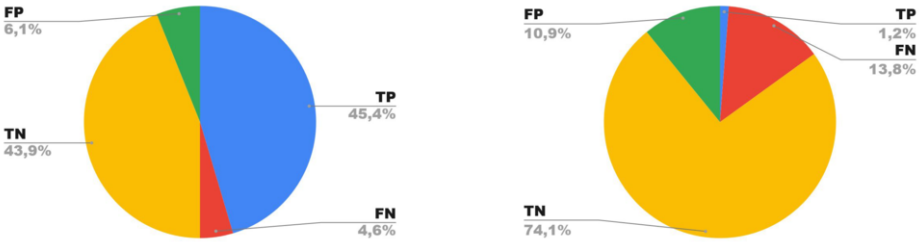
To test the utility of this model, a new test is defined. A smaller dataset created with the data of 20.000 customers is used, and partitionig it into 70% for training and 30% for test. Again 10 executions were done and saving the average scores. This dataset was used with the original model (adapting the time window to the oblivion model equivalence) and the results were compared with the new method applying transformation for several time dependent variables: first only the errors, then the errors and time spent by the customer in the platform, and finally for all the time dependent metrics. Additionally, the models were tested also applying PCA.

The results of figure 4 show that the results are better by applying this new oblivion model.

5. Results

The final model was developed using the HGBC algorithm. Their parametres were estimated using random search. The dataset used to develop the model was the dataset de-

scribed in section 2, and the transformation of the oblivion model was applied (with a d of 30 days) and making a PCA with the resulting metrics. The partition defined was 60% of the dataset to train, 30% to test and 10% to validate. The model was trained 10 different times and calculating the score average. To manage the imbalance problem oversampling was used in the test and training sets, but no with the validation set in order to calculate the outperformance of the model over a real distribution of the data.



(a) On the test set (b) On the validation set.

Figure 5. Results of the final model.

Several thresholds to decide if an instance is negative or positive have been tested, resulting that 50% worked better. The final score of the model was 86% in the test set and 90% for the validation data, as we can see in figure 5. In this model the negatives are the no churned customers and the positives are the customer that churned.

6. Conclusions

In this work, a comparative study of algorithms for predicting the customer churn in the Pay-TV sector has been done. Oversampling and undersampling methods were tested for handing the class imbalance problem inherent in this problem. A new model considering the fact that people forget things happened long time ago is presented and named the oblivion model. This model improves the results got without applying it because the use of metrics that are time dependent. Finally, model wich can discriminate the churn of customers is constructed and presented.

7. Acknowledgments

This project has been funded by the Ministry of Economy and Commerce with project contract TIN2016-88835-RET and by the Universitat Jaume I with project contract UJI-B2020-15, and by Mirada.TV.

References

[1] V Umayaparvathi and K Iyakutti. A survey on customer churn prediction in telecom industry: Datasets, methods and metrics. *International Research Journal of Engineering and Technology (IRJET)*, 3(04), 2016.

[2] Ya Gao, Guangquan Zhang, Jie Lu, and Jun Ma. A bi-level decision model for customer churn analysis. *Computational Intelligence*, 30(3):583–599, 2014.

[3] Xiaobing Yu, Shunsheng Guo, Jun Guo, and Xiaorong Huang. An extended support vector machine forecasting framework for customer churn in e-commerce. *Expert Systems with Applications*, 38(3):1425–1430, 2011.

[4] Koen W De Bock and Dirk Van den Poel. An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction. *Expert Systems with Applications*, 38(10):12293–12301, 2011.

[5] Jonathan Burez and Dirk Van den Poel. Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3):4626–4636, 2009.

[6] Xinjian Guo, Yilong Yin, Cailing Dong, Gongping Yang, and Guangtong Zhou. On the class imbalance problem. In *2008 Fourth international conference on natural computation*, volume 4, pages 192–201. IEEE, 2008.

[7] V Umayaparvathi and K Iyakutti. Applications of data mining techniques in telecom churn prediction. *International Journal of Computer Applications*, 42(20):5–9, 2012.

[8] Mohammed Hassouna, Ali Tarhini, Tariq Elyas, and Mohammad Saeed AbouTrab. Customer churn in mobile markets a comparison of techniques. *arXiv preprint arXiv:1607.07792*, 2016.

[9] David L García, Àngela Nebot, and Alfredo Vellido. Intelligent data analysis approaches to churn as a business problem: a survey. *Knowledge and Information Systems*, 51(3):719–774, 2017.

[10] Jaehyun Ahn, Junsik Hwang, Doyoung Kim, Hyukgeun Choi, and Shinjin Kang. A survey on churn analysis in various business domains. *IEEE Access*, 8:220816–220839, 2020.

[11] Adnan Amin, Feras Al-Obeidat, Babar Shah, May Al Tae, Changez Khan, Hamood Ur Rehman Durrani, and Sajid Anwar. Just-in-time customer churn prediction in the telecommunication sector. *The Journal of Supercomputing*, 76(6):3924–3948, 2020.

[12] John Hadden, Ashutosh Tiwari, Rajkumar Roy, and Dymitr Ruta. Computer assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research*, 34(10):2902–2917, 2007.

[13] Joffrey L Leevy, Taghi M Khoshgoftaar, Richard A Bauder, and Naeem Seliya. A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5(1):1–30, 2018.

[14] Yongbin Zhang, Ronghua Liang, Yeli Li, Yanying Zheng, and Michael Berry. Behavior-based telecommunication churn prediction with neural network approach. In *2011 International Symposium on Computer Science and Society*, pages 307–310. IEEE, 2011.

[15] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[16] Vicente López Oliva. Predicción de churn en una plataforma de pay-tv mediante machine learning. Technical report, Universitat Jaume I, 2020.