International Journal of Engineering and Information Systems (IJEAIS)
ISSN: 2643-640X
Vol. 6 Issue 11, November - 2022, Pages: 1-7

Performing Customer Churn Analysis and Classification Using Machine Learning

DA Ajiputra, M Arfan, M Somantri

Department of Electrical Engineering, Undip, Central Java, Indonesia
arfan.undip@gmail.com

***Abstract**: Customer Churn Analysis and Classification with the ultimate goal of finding out what factors cause customer churn and how to overcome them. The methodology used in this research project is the Cross Industry Standard Process for Data Mining or CRISP-DM. method, with four classification models, namely: Logistic Regression, Decision Tree, Randomforest without Hyperparameter Tuning, and Random Forest with Hyperparameter Tuning. the author takes the learning path 'Accelerated Machine Learning Program. Customer Churn Analysis and Classification with the ultimate goal of finding out what factors cause customer churn and how to overcome them. The methodology used in this research is the Cross Industry Standard Process for Data Mining or CRISP-DM. method, with four classification models, namely: Logistic Regression, Decision Tree, Randomforest without Hyperparameter Tuning, and Random Forest with Hyperparameter Tuning. The research project that the author is doing is Customer Churn Analysis and Classification with the ultimate goal of finding out what factors cause customer churn and how to overcome them. The methodology used in research is the Cross Industry Standard Process for Data Mining or CRISP-DM. method, with four classification models, namely: Logistic Regression, Decision Tree, Randomforest without Hyperparameter Tuning, and Random Forest with Hyperparameter Tuning. Customer Churn Analysis and Classification with the ultimate goal of finding out what factors cause customer churn and how to overcome them. The methodology used in this research is the Cross Industry Standard Process for Data Mining or CRISP-DM method, with four classification models, namely: Logistic Regression, Decision Tree, Randomforest without Hyperparameter Tuning, and Random Forest with Hyperparameter Tuning.*

Keywords— machine learning; random forest; CRISP-DM

1. INTRODUCTION

The rate at which customers decide not to buy more of a company's goods or services is known as customer churn, also known as attrition. An approach to calculating this rate is through customer attrition analysis. Churn analysis essentially shows you what proportion of your consumers don't make a return purchase compared to the proportion that do repeat business. You might be able to see trends that can prevent failure or propel an already successful product or service to the next level by delving further into these numbers.

The financial effects of consumers leaving are also measured by high-performing businesses, who then compare those results to key performance indicators (KPIs) important to the profitability of the company. This KPI can be calculated over different timeframes and its results can be trended by Machine Learning.

Machine Learning is a branch of science that is part of artificial intelligence (artificial intelligence), with programming that allows computers to be intelligent, behave like humans, and automatically expand their understanding through experience. Machine learning focuses on developing systems that can learn to make decisions independently without the need for repetitive human programming. This allows machines not only to behave when making decisions, but also to adapt to the ongoing transformation. Machine learning works when there is data available as input to analyze large amounts of data (big data) to recognize certain patterns. Data is the input material used to implement learning (training) so that the machine can produce the appropriate analysis. Machine Learning includes training data and testing data. The training data is used to train the machine learning algorithm, while the test data is used to determine the performance of the machine learning algorithm that has been trained to recognize new data that is not specified in the training data.

2. METHODOLOGY

2.1 Costumer Churn

Customer churn is a condition in the business world that shows that customers stop using the services of a company or customers who switch services to other companies. This becomes important for companies that do business with a subscription model, such as insurance, telecommunications, or banking [2]. In addition, with the increasing number of customer churn, the value of revenue from the company will also decrease. Therefore, it is important to understand customer churn so that it can be immediately addressed by the company, and another reason is that the cost of getting new customers can be more expensive than dealing with customer churn [3].

2.2 Random Forest

Transformer oil is a complex mixture of hydrocarbon molecules. Mineral insulating oil is formed from several molecules containing the chemical groups CH3, CH2, and CH linked by carbon molecular bonds.

Logistic Regression is the most important procedure used to model binary variables (0, 1) based on one or more other variables, called predictors [4]. The modeled binary variable is generally referred to as the response variable, or the dependent variable [5]. The independent variable can be nominal level, ordinal (rating), interval, or scale (or continuous data). This can then be extended to predictions using a set of predictors on a set of dependent variables. With logistic regression, a model is developed from the statistic that best describes the relationship between variables [6].

Decision Tree is an algorithm consisting of nodes (root, branch, leaf) and edges. In general, this algorithm has two processes, namely composing a decision tree and making rules, and calculating entropy to select the attribute that has the highest gain value [7]. Each node in the decision tree represents a characteristic (attribute), each connection (branch) represents a decision (rule) and each leaf represents an outcome (categorical or continuous). Because decision trees mimic human-level thinking, it is very easy to take data and make it good for interpretation. The concept is to create a tree like this for all data and process the results in each leaf [8].

Random forest is a bagging method, which is a method that generates a number of trees from sample data where the creation of one tree during training does not depend on the previous tree then decisions are taken based on the most voting. [9]. Random forest modeling can be used to perform classification and regression. In doing random forest modeling, it is important to perform hyperparameter tuning [10].

2.3 CRISP-DM

The Cross Industry Standard Process for Data Mining (CRISP-DM) methodology is a general problem-solving strategy from business as a research flow consisting of six stages as follows [4].

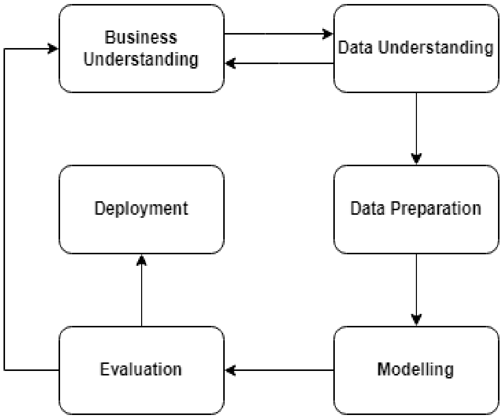


Fig. 1. CRISP-DM Methodology

Table 1: CRISP-DM methodological process

Phase	Explanation
Business Understanding	Steps to understand business process with Formulate part problems to be solved that can be seen from the business situation of an industry in order to achieve goals.
Data Understanding	The data understanding stage begins with the collection of initial information and continues with activities that allow for initial adjustment to the data, identify data quality problems, gain initial understanding of the data, and/or uncover interesting subsets to form hypotheses such as hidden information.

Data Preparation	A process for preparing the final dataset which will later be used in the modeling phase from the initial raw dataset.
Modelling	The data modeling stage consists off election technique. modeling, building test cases and models. To build a model, certain parameters must be set. To assess the models, it is appropriate to evaluate the model against the evaluation criteria and choose the best one.
Evaluation	The evaluation stage of the accuracy of the results that have been obtained in the modeling process is to determine the performance of the algorithm used.
Deployment	The process of presenting the information obtained in a guide. Guides can be reports, databases, and web pages

3. RESULT

3.1 Data Analysis

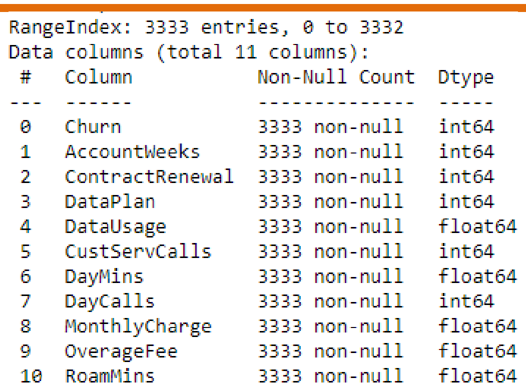
The data used is a dataset that is determined on the topic of the final project, namely a dataset that is accessed from the Kaggle website (kaggle.com) with the dataset name "Customer Churn". This dataset is a dataset regarding customer churn in a telecommunications company. Insight exploration is carried out using the CRISP-DM methodology.

Goals of this final project is to see what factors influence customer churn, and what efforts can be made to reduce churn. The algorithm used is a classification algorithm. The dataset has dimensions of 11 columns x 3333 rows. The explanation of each column is as follows.

Table 2*:* Columns in the dataset

Columns	**Desc**	**Data Type**	**Fill Data**
Churn	Whether customer churn or not	Category	1 = Customer Churn 0 = Not
Account Weeks	Weekly number of customers who have active accounts	Numerical	
ContractRenewal	Has the customer recently renewed the contract	Category	1 = Yes 0 = Not
DataPlan	Does the customer have a data package	Category	1 = Yes 0 = Not
DataUsage	User data	Numerical	
CustServCalls	Number of calls	Numerical	

Before conducting exploration, it is necessary to check the data first. The checks carried out are regarding missing values, and outliers. However, outliers can be done when exploring numerical data.



```
RangeIndex: 3333 entries, 0 to 3332
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   Churn            3333 non-null   int64  
 1   AccountWeeks     3333 non-null   int64  
 2   ContractRenewal  3333 non-null   int64  
 3   DataPlan         3333 non-null   int64  
 4   DataUsage        3333 non-null   float64
 5   CustServCalls    3333 non-null   int64  
 6   DayMins          3333 non-null   float64
 7   DayCalls         3333 non-null   int64  
 8   MonthlyCharge    3333 non-null   float64
 9   OverageFee       3333 non-null   float64
 10  RoamMins         3333 non-null   float64
```

Fig. 2. Checking for missing values from the dataset

From Figure 2, after checking for missing values, it shows that the dataset does not have a missing value. Because there is no missing value, it is not necessary to take action to overcome the missing value.

Furthermore, because the purpose of this project is customer churn, the target column to be used is the churn column. For comparison between customers who churn and those who do not from the churn column, can be seen in the following picture.

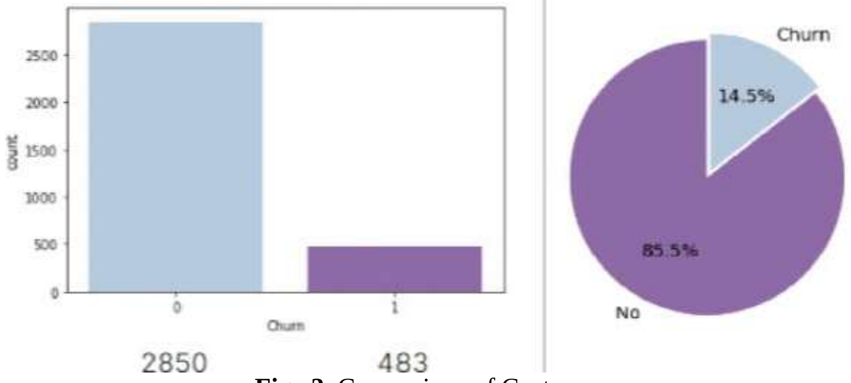


Fig. 3. Comparison of Costumer

After looking at the comparison of customers who churn and those who don't, the next step is to look at the correlation of each column. To see the correlation between columns, the feature used is heatmap. The result of the heatmap is as shown below.

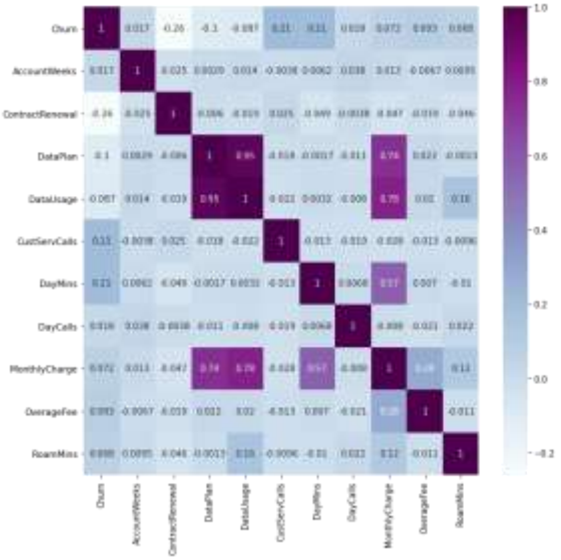


Fig. 4. Correlation between columns with heatmap

From the heatmap, if the correlation between the 2 columns is higher (closer to 1) it will be purple, and the weaker the correlation will be blue. Referring to this, the 'Churn' column does not have a strong correlation with other columns so that further exploration of the dataset is necessary. In addition, from the heatmap there are insights that can be used for future processes. The insights obtained are:

- There is a very strong correlation between the columns 'DataUsage' and the 'DataPlan' column.
- The next strongest correlation is between Data Usage ('DataPlan', 'DataUsage', 'DayMins') and 'MonthlyCharge'.

In exploring further to see what column affects churn, the dataset can be divided into 2 categories based on the column data type, namely categorical data, and numerical data. For categorical data, it can be seen the relationship between column 'Churn' and column 'DataPlan', and 'ContractRenewal'. To see the relationship, one of the functions that can be used is the crosstab function. The result of the relationship between 'Churn' and 'ContractRenewal' with the crosstab function is as follows.

3.2 Data Preparation

This stage is the stage of preparing the dataset before doing the modeling. The dataset has to do some preparation, including overcoming the imbalance problem, changing the dataset which has object or string/char data types, and doing the separation for train and test.

To overcome the imbalance problem. The balance problem is to equalize the value of the target column, in this case it is the 'Churn' column. Overcoming the imbalance problem will train the program to look at the problem the same way. If not addressed, because the number of customers who churn and who do not differ very much, the program will tend to make judgments towards not churn. To overcome the imbalance problem can be done by using undersampling or upper sampling. Undersampling is equating the target data value with its minimum value, while Oversampling is equating the target data with its maximum value. In the case of this final project, the undersampling value is 483, while the Oversampling is 2850. The data is taken from churn values and those that do not refer to or refer to Figure 2.

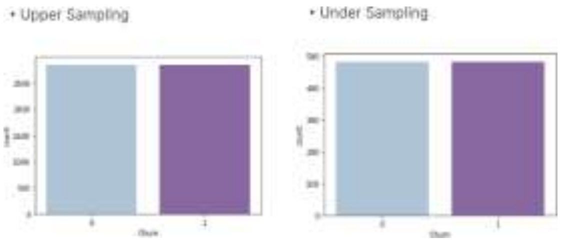


Fig. 5. Oversampling and undersampling of data

Both Oversampling and undersampling data will later be used for comparison in the final results.

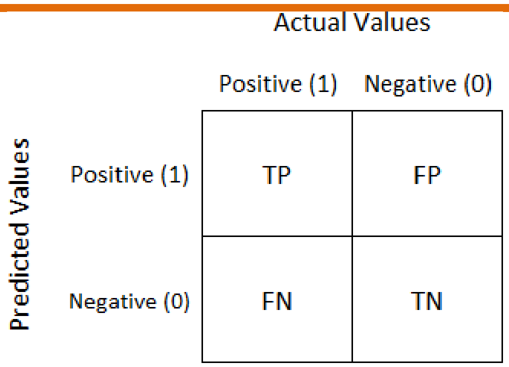
After overcoming the imbalance problem, the next step is to change object data or string/character data types to numeric data types. This data conversion is commonly referred to as 'One-Hot Encoding'. In doing one-hot encoding, you can use the 'pd.get_dummies' function which will convert the data automatically. However, because the dataset used does not have object type data, this does not need to be done.

Next is the trains test split or the separation of the dataset into data to train the program and test the program. Where this separation is done automatically. The share between train and test is 60% for train and 40%.

3.3 Modelling and Evaluation

Modeling carried out in four classification models, namely: Logistics Regression, Decision Tree, Random Forest without Hyperparameter Tuning, and Random Forest with Hyperparameter Tuning.

At this stage, the quality of the modeling used will be evaluated. The evaluation is done by looking at the confusion matrix and also looking at the accuracy of the classification report. The confusion matrix will categorize the classification results into four types automatically from the program, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Where a value of True Positive means the program detects will be positive for the target (in the case of the project this ending is churn) and true is positive. A False Positive value means that the program detects that it will be positive on the target, but the result is false. A False Negative value means that the program detects a negative value against the target, but the result is false. Finally, a True Negative value means that the program detects that it will be negative to the target, and true is negative.



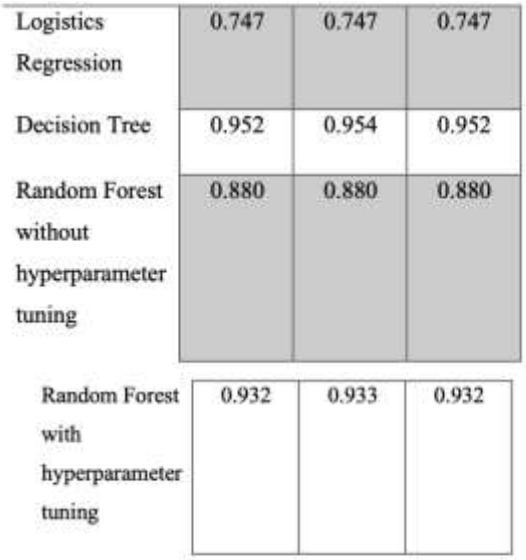
		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig. 6. Confusion Matrix

Meanwhile, the classification report is a table that contains the values of accuracy, precision, and recall.
The evaluation is done by using data with Oversampling and undersampling. The results of the testing are as follows.

1. Oversampling data

Table 3: Classification Report for Oversampling data

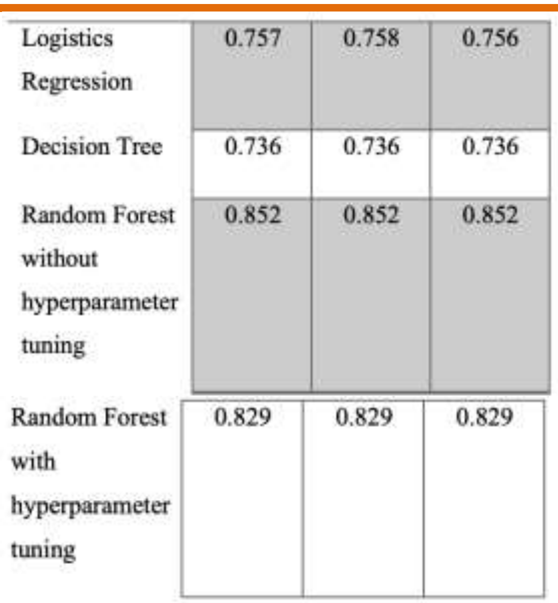


Logistics Regression	0.747	0.747	0.747
Decision Tree	0.952	0.954	0.952
Random Forest without hyperparameter tuning	0.880	0.880	0.880
Random Forest with hyperparameter tuning	0.932	0.933	0.932

From the confusion matrix and the accuracy of the classification report, the best modeling for Oversampling data is the decision tree with an accuracy of 0.952 or 95.2%.

2. Undersampling data

Table 4: Classification Report for Undersampling data



Logistics Regression	0.757	0.758	0.756
Decision Tree	0.736	0.736	0.736
Random Forest without hyperparameter tuning	0.852	0.852	0.852
Random Forest with hyperparameter tuning	0.829	0.829	0.829

From the confusion matrix and the accuracy of the classification report, the best modeling for undersampling data is random forest without hyperparameter tuning with an accuracy of 0.852 or 85.2%.

4. Condlusion and Suggestion

Considering the accuracy of the confusion matrix and classification report, the best modeling for oversampled data is a decision tree with an accuracy of 0.952 or 95.2%. .Beside that the best modeling for under sampled data is a random forest without hyperparameter tuning, with an accuracy of 0.852 or 85.2%.

The more Customer Service Calls, the more vulnerable it is to Churn. Therefore, must immediately optimize services so that there are no more disturbances that will lead to complaints from customers. Optimize call time pricing for customer segment. Introduce data packages to customers who use data without data packages as soon as possible. Optimize the price of data packages for customer's segment

Considering the accuracy of the confusion matrix and classification report, the best modeling for oversampled data is a decision tree with an accuracy of 0.952 or 95.2%.

Make Exploratory Data Analysis more interesting. Try modeling with other models, and again compare all the modeling results so that the best modeling will be obtained

5. References

[1] Retnoningsih, E. and Pramudita, R., 2020. Understanding Machine Learning With Supervised And Unsupervised Learning Techniques Using Python. Bina Insani ICT Journal, 7(2), pp.156-165.

[2] elik, O. and Osmanoglu, UO, 2019. Comparing to techniques used in customer churn analysis. Journal of Multidisciplinary Developments, 4(1), pp.30-38.

[3] Osman, Y. and Ghaffari, B., 2021. Customer churn prediction using machine learning: A study in the B2B subscription based service context. Sweden: DiVA

[4] Pérez López, C. 2021. Data Mining. The CRISP-DM Methodology. The CLEM Language and IBM SPSS Modeler. United Kingdom: Lulu.com.

[5] Hilbe, JM, 2016. Practical guide to logistic regression. Florida: CRC Press.

[6] Connelly, L., 2020. Logistic regression. Medsurg Nursing, 29(5), pp. 353-354.

[7] Budiman, E., Kridalaksana, AH and Wati, M. 2017. Performance of decision tree C4. 5 algorithm in student academic evaluation. International Conference on Computational Science and Technology, pp. 380-389. Singapore: Springer.

[8] Patel, HH and Prajapati, P., 2018. Study and analysis of decision tree based classification algorithms. International Journal of Computer Sciences and Engineering, 6(10), pp.74-78.

[9] Wibowo, AT, 2016. Implementation of SPAM Detection Algorithm with Image Information Inserted with SVM and Random Forest Methods, Doctoral dissertation, Sepuluh Nopember Institute of Technology, Surabaya.

[10] Cheng, L., et al., 2019. Applying a random forest method approach to the travel mode choice behavior model. Travel behavior and society, 14, pp.1-10