Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 19, no. 10, October 2023, pp. 663-675
DOI: 10.23940/ijpe.23.10.p4.663675

Customer Churn Analysis using Spark and Hadoop

Priyanshu Verma, Ishan Sharma, Sonia Deshmukh, and Rohit Vashisht*

KIET group of Institutions, Dr. A.P.J. Abdul Kalam Technical University, Lucknow, India

Abstract

Predicting Customer churn is one of the telecommunication industry's biggest challenges. Why did their customers quit using their product, site, service, or subscription? Machine learning with Spark and Hadoop has considerably increased the ability to predict customer behaviours. The most popular predictive models, such as logistic regression, Binary Classification Evaluator, and Multi Classification Evaluator, have been used in the prediction process. Enhancing and outfit approaches are used on the training dataset to examine the impact on model effectiveness. Additionally, to further optimize the hyperparameters and produce the models, a K-fold cross-validation method is utilized to train the dataset. Finally, the test data were examined by the AUC-ROC curve and confusion matrix. In this research, an adaptation of Spark and Hadoop frameworks is made to predict customer churn. The data is pre-processed, feature analyses are performed, and the feature selection is carried out using the Vector Assembler algorithm. This study aims to analyse customer behaviors by using a dataset.

Keywords: Hadoop and Spark; Machine Learning; Logistic regression; Random Forest; Vector Assembler; Binary Classification Evaluator

1. Introduction

Churn prediction is the process of forecasting a customer's preference and behaviour to stop using a product, website, or service. In recent years, it has developed into a highly disputed topic of research. This study uses a dataset of telecommunications customers to predict the chance of customer attrition. The increased popularity of big data analysis technologies has resulted in their broad use in a variety of telecom enterprises. The fundamental difficulty that business owners confront is automating their processes and analysing massive amounts of data more precisely and accurately. Dealing with the constant influx of data from many sources presents a big challenge. The capacity to predict client attrition based on data and interactions is a critical requirement. This requirement is heightened for organizations with a large customer base, such as telecommunications corporations. A comparative examination of customer data is performed in this paper. Apache Spark has grown in popularity because of its remarkable performance and efficiency in processing enormous amounts of data faster than Apache Hadoop. Notably, it addresses MapReduce's constraints. The ML package, which is a higher-level API built on datasets, has superseded the RDD-reliant Spark machine learning APIs in the MLlib package. This change allows for more realistic ML pipelines, especially for feature conversions. This study aims to examine the accuracy, model training, and model evaluation of these two programs. It clarifies the practical differences between the two packages and distinguishes their advantages and disadvantages based on the results produced by processing the same dataset with the same method.

The structure of this paper is as follows: Section 2 offers a comprehensive review of the pertinent literature. Section 3 delves into the dataset employed in this study and outlines the steps taken to analyse the extensive data it details. It further elaborates on the chosen algorithm and the reasons behind its selection for this investigation. The fourth section centres around the evaluation and discoveries, while the conclusion concisely summarizes the research results, including a description of the system's architecture in (Figure 1).

* Corresponding author.
E-mail address: rohit.vashisht@kiet.edu

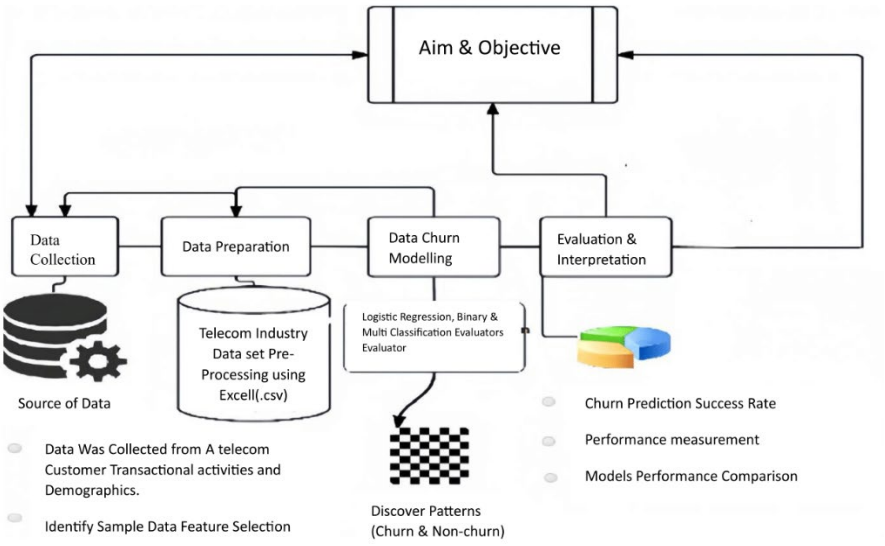


Figure 1. Customer Churn Prediction Flowchart

2. Related Work

Within literature, numerous techniques have been devised to forecast churn in telecom companies as well as other industries. Machine learning and data mining have been predominantly employed in these strategies. While some researchers concentrate on utilizing a single data mining technique to extract information and behaviours from consumers, others are dedicated to comparing different methods for customer churn predictions.

Gavril et al. Implemented a state-of-the-art data mining algorithm to forecast the churn among prepaid customers, leveraging a dataset comprising of call information from 3333 consumers with 21 distinct attributes. The dependent churn parameter was binary with two possible values: yes or no. Notable features encompassed information pertaining to the number of received and sent messages, along with voicemail usage for each individual client.

Idris et al. created an approach that employed AdaBoost and genetic programming to simulate churn within the telecommunications industry. The model underwent evaluation using two widely used datasets. The first dataset originated from Orange Telecom, while the second was provided by cell2cell. Notably, the cell2cell dataset exhibited an impressive accuracy rate of 89 percent, whereas the other dataset achieved a comparatively lower accuracy of 63 percent.

Huang et al. [1] conducted a study on the challenge of client attrition within the context of a big data platform. The researchers have demonstrated that when taking into account elements like data volume, complexity, and mobility, big data has a significant impact on improving churn prediction. A big data platform was used as a solution to address the data management issues that China's leading telecoms business encountered when handling data from the Procedure Assist and Business Support departments. The AUC statistic was used to assess the Random Forest technique.

Burez and Van den Poe et al. [2] examined the challenge posed by imbalanced datasets in prediction models and even assessed the efficacy of various techniques, including gradient boosting model, random sampling, advanced under-sampling, and weighted random forests. The evaluation of the given model was conducted using metrics such as AUC and Lift. The results indicated that the under-sampling strategy outperformed the other analysed techniques.

J. Burez et al. [3] performed a study with the goal of addressing the issue of class inequality by employing random forest and logistic regression techniques, as well as resampling methods. The analysis also involved the use of boosting methods, with performance evaluation based on metrics such as AUC and Lift. The researchers investigated the impact of reducing sampling methods, including CUBE, but found no substantial improvement in efficiency. However, it was determined that optimization-based sampling strategies offer a more effective approach to tackling the issue of class imbalance.

J. Hadden et al. [4] provided a thorough analysis of various feature selection techniques and evaluated multiple machine learning models. It was observed that decision trees outperformed other models in prediction accuracy. The authors

emphasized the importance of optimization strategies in feature selection as they contribute to improved prediction algorithms. After summarizing previous methods, the authors suggested future directions for scientific research in the field.

P. Kisioglu et al. [5] employed Bayesian belief networks (BBN) to predict customer attrition. Correlation analysis and cross-tests were conducted as part of the investigation. The findings revealed that BBN proved to be a highly effective approach for churn prediction. In addition, the authors put forward suggestions for future research directions in this area.

K Coussement et al. [6] addressed the challenge of churning predictions by customers using support vector machines (SVM), logistic regression (LR), and random forest (RF) techniques. Initially, SVM exhibited comparable performance to LR and RF. However, upon considering optimal parameter selection, SVM outperformed both LR and RF in terms of PCC (Pearson correlation coefficient) and AUC (area under the curve).

In this paper, numerous strategies have been employed, predominantly cantered around the utilization of machine learning and data mining techniques. While some researchers focus on extracting information and understanding consumer behaviour using a single data mining technique, others take a comparative approach, examining various methods for churn prediction. In the process of feature development for machine learning algorithms, the feature engineering process is duly considered. To prepare the data and assess performance, a big data platform was employed, and the outcomes of 4 different algorithms for machine learning were compared.

3. Methodology

To provide a suitable model for predicting and analysing customers' churn, by using machine learning tools with Spark and Hadoop, a methodology is designed as below (Figure 1).

3.1 Data Pre-processing

Data pre-processing is a crucial approach for transforming raw data into a format that machine learning algorithms can use. It serves as the initial step in employing machine learning algorithms. During machine learning research, it is imperative to cleanse and structure the data, and this cleaning and formatting process is mandatory for any data manipulation. Real-world data contains noise, missing values, or is presented in a disorderly manner, hindering direct use with machine learning models. Data pre-processing becomes necessary to clean and organize the data, thereby preparing it for effective utilization by a machine learning model. This process ultimately enhances the model's accuracy and efficiency. The foremost requirement for developing a machine learning model is a dataset, as the model's functionality entirely relies on the provided data. A dataset is essentially a collection of data formatted specifically for a given task. For example, when constructing a machine learning model for commercial purposes, the dataset required would differ from the data necessary for analysing liver patients. Consequently, each dataset holds its own distinct characteristics. Typically, the dataset is saved in a CSV file before being utilized in the code. During the research process, CSV formatted data was employed.

3.2 Feature Selection

The characteristics within the data utilized in machine learning models are referred to as features. Each column in the dataset represents a feature. To train a predictive model effectively, it is essential to include only the relevant features. If there are an excessive number of features, the model may capture irrelevant patterns and learn from noise. Feature selection is the procedure of identifying and selecting the crucial parameters within our data (Figure 2).

Figure 2. Feature Selection on data

3.3 Testing and Training

The testing and training method is the most crucial component impacting machine learning performance. A good training method raises the quality of the established system. Datasets are divided into two portions for training and testing done by researchers. However, the separation procedure follows strict guidelines. The division of training and testing data is a crucial factor in determining the success of a model. When there is a strong correlation between the features and labels, a common approach is to split the data into a training set and a test set using an 80 percent to 20 percent ratio. This means that 80 percent of the data is used for training the model, while the remaining 20 percent is reserved for testing and evaluating its performance. (Figures 3 and 4).

Figure 3. Training and Test Data

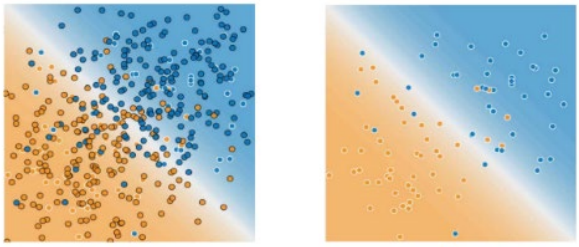


Figure 4. Testing and Training phase

3.4 Machine Learning Algorithms

Machine learning (ML), which falls under the umbrella of artificial intelligence (AI), empowers computer systems to make more accurate predictions without being explicitly programmed to do so. By utilizing past data as input, machine learning algorithms can forecast new output values. Recommendation systems are one example of machine learning applications, and they are frequently used in various domains such as malware threat detection, spam filtering, fraud detection, business process automation (BPA), and predictive maintenance, among others. The significance of machine learning lies in its ability to help organizations identify customer behaviour trends, operational patterns, and facilitate the development of new products. Machine learning plays a pivotal role in the operations of numerous highly successful companies today, including Facebook, Google, and Uber. The research study incorporates several algorithms in its analysis.

3.4.1 Vector Assembler

- The VectorAssembler is a transformer that consolidates a given set of columns into a single vector column.
- It is particularly useful in training machine learning models for example logistic regression and decision trees, as it combines raw features with features generated by different feature transformers into a unified feature vector.
- The VectorAssembler accepts input column types that include integer types, boolean types, and vector types. The values of each row's input columns will be concatenated in the specified order to form a vector [7].

3.4.2 Logistic Regression

A supervised machine learning model that goes under the categorization type is PySpark Logistic Regression. This algorithm defines the connections between the data and classifies the data into categories based on those connections. Logistic regression is the primary classification approach to the behaviour and is simpler to compute. It is based on the training and testing of the data model by PySpark's machine learning model. The dependent variable's probability is anticipated. It is a predictive analysis describing variables' relationships and characterizes data. This assignment will investigate the various Logistic Regression approaches [8].

3.5 Machine Learning Classification

3.5.1 BinaryClassificationEvaluator

BinaryClassificationEvaluate extends Evaluator and implements Default Params Writable.

- The Evaluator is designed for binary classification and is capable of processing a raw prediction along with two input columns for labelling purposes.
- The raw prediction column can contain either a binary 0/1 prediction, the probability of label 1, or a vector type (a vector of length-2 comprising raw predictions, scores, or label probabilities) [9].

3.5.2 MultiClassificationEvaluator

In machine learning, the classification process revolves around assigning each new data point to a specific category from a predefined set of categories. This assignment is achieved through a mapping function that is created using a training set of data, where the category membership of each observation is already known. The three widely used types of classifiers are as follows:

- Binary: This type of classification involves two distinct and exclusive outcomes, such as determining whether something is classified as “Hotdog” or “Not.”
- Multiple classes: In this classification scenario, there are multiple mutually exclusive outcomes, such as classifying items into categories.
- Multi-label: This classification deals with cases where there can be multiple overlapping possible outcomes. For example, a research paper may contain information related to politics, business, and sports simultaneously.

3.6 Classification Report / Confusion Matrix

The table serves the purpose of categorizing issues to identify the source of errors in the model. The rows correspond to the actual classes for which the recommendations were made. The forecasts they made are displayed in the columns. This table makes it easy to see whose forecasts were incorrect or correct. Confusion matrices may come from logistic regression predictions. For now, we’ll use NumPy to generate real and anticipated numbers [10].

4. Experimentation

4.1 Dataset

In the telecom industry, various types of data are employed to construct churn models. However, the preferred dataset format for such models is CSV. The dataset comprises of details such as company name, etc., and encompasses information about customers’ total purchase and subscription. The customer data encompasses all relevant information pertaining to customers, including their services, contact details, and subscribed services. Additionally, it also includes information regarding the customer’s location, and the name of their company that is linked with telecom industries. The dataset is used to predict the accuracy and prediction (Figures 5 and 6) [11].

	A	B	C	D	E	F	G	H	I	J	K
1	Names	Age	Total_Purchase	Account_Manager	Years	Num_Sites	Onboard_date	Location	Company	Churn	
2	Cameron Williams	42	11066.8	0	7.22	8	30-08-2013 07:00	10265 Elizabeth Mission Barkerbu	Harvey LLC	1	
3	Kevin Mueller	41	11916.22	0	6.5	11	13-08-2013 00:38	6157 Frank Gardens Suite 019 Ca	Wilson PLC	1	
4	Eric Lozano	38	12884.75	0	6.67	12	29-06-2016 06:20	1331 Keith Court Alyssahaven, DE	Miller, Johnson and Wallace	1	
5	Phillip White	42	8010.76	0	6.71	10	22-04-2014 12:43	13120 Daniel Mount Angelabury,	Smith Inc	1	
6	Cynthia Norton	37	9191.58	0	5.56	9	19-01-2016 15:31	765 Tricia Row Karenshire, MH 7	Love-Jones	1	
7	Jessica Williams	48	10356.02	0	5.12	8	03-03-2009 23:13	6187 Olson Mountains East Vince	Kelly-Warren	1	
8	Eric Butler	44	11331.58	1	5.23	11	05-12-2016 03:35	4846 Savannah Road West Justin	Reynolds-Sheppard	1	
9	Zachary Walsh	32	9885.12	1	6.92	9	09-03-2006 14:50	25271 Roy Expressway Suite 147	Singh-Cole	1	
10	Ashlee Carr	43	14062.6	1	5.46	11	29-09-2011 05:47	3725 Caroline Stravenue South Cl	Lopez PLC	1	
11	Jennifer Lynch	40	8066.94	1	7.11	11	28-03-2006 15:42	363 Sandra Lodge Suite 144 Soutl	Reed-Martinez	1	
12	Paula Harris	30	11575.37	1	5.22	8	13-11-2016 13:13	Unit 8120 Box 9160 DPO AA 4343	Briggs, Lamb and Mathews	1	
13	Bruce Phillips	45	8771.02	1	6.64	11	28-05-2015 12:14	Unit 1895 Box 0949 DPO AA 4024	Figueroa-Maynard	1	
14	Craig Garner	45	8988.67	1	4.84	11	16-02-2011 08:10	897 Kelley Overpass Suite 349 W	Abbott-Thompson	1	
15	Nicole Olson	40	8283.32	1	5.1	13	22-11-2012 05:35	11488 Weaver Cape Hernandezb	Smith, Kim and Marshall	1	
16	Harold Griffin	41	6569.87	1	4.3	11	28-03-2015 02:13	1774 Peter Row Apt. 712 New Au	Snyder, Lee and Morris	1	
17	James Wright	38	10494.82	1	6.81	12	22-07-2015 08:38	45408 David Path East Kimberlysl	Sanders-Pierce	1	
18	Doris Wilkins	45	8213.41	1	7.35	11	03-09-2006 06:13	28216 Wright Mount Apt. 356 Ali	Andrews, Adams and Davis	1	
19	Katherine Carpenter	43	11226.88	0	8.08	12	22-10-2006 04:42	Unit 4948 Box 4814 DPO AP 4266	Morgan, Phillips and Harrell	1	
20	Lindsay Martin	53	5515.09	0	6.85	8	07-10-2015 00:27	69203 Crosby Divide Apt. 878 Par	Villanueva LLC	1	
21	Kathy Curry	46	8046.4	1	5.69	8	06-11-2014 23:47	9569 Caldwell Crescent Tanyabo	Berry, Orr and Cabrera	1	
22	Dean Miller	41	9771.22	0	5.81	11	30-05-2013 00:42	803 Kelli Crossing Apt. 169 Jimen	Parks-Bradley	1	
23	Kevin Ramos	56	12217.95	1	5.79	11	17-11-2016 14:37	5775 Jared Stream Apt. 881 Port	Olsen LLC	1	
24	Jennifer Wood	35	9381.12	1	6.78	11	27-03-2006 20:52	1493 Phillips Haven Lake William,	Clark, Campbell and Armstrong	1	
25	Paul Walker	41	10474.94	0	6.4	12	02-01-2012 05:08	73006 Patty Avenue Apt. 646 Ne	Dalton LLC	1	
26	Lindsey Day	55	11158.5	1	4.86	10	07-01-2007 01:21	85934 Dakota Mall Timothyville,	Thompson, Hansen and Sanchez	1	

Figure 5. Dataset

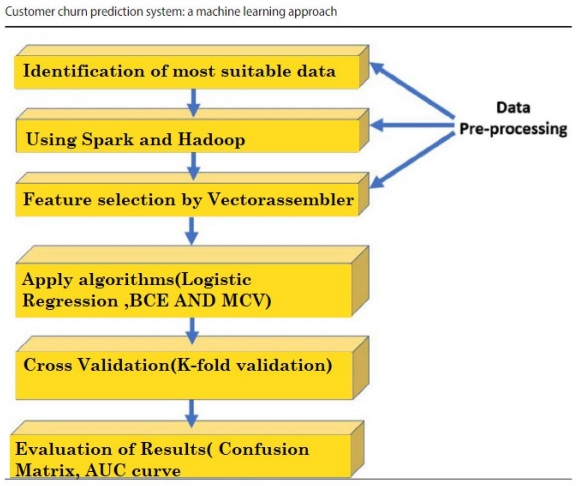


Figure 6. Dataset pre-processing

4.2 Big Data Platform

4.2.1 Hadoop

Apache Hadoop is a platform, available as open source, designed for the storage and processing of extremely large datasets, ranging from gigabytes to petabytes in size. Rather than relying on a single massive computer, Hadoop enables the clustering of multiple computers, allowing for more efficient parallel analysis of extensive datasets. Hadoop consists of four main components [12].

- Hadoop Distributed File System (HDFS) – HDFS, which is known to be a Hadoop Distributed File System, facilitates data access across any Hadoop cluster. A cluster refers to a collection of interconnected computers. HDFS, along with other Hadoop-related technologies, plays a vital role in the management and analysis of massive data volumes, reaching scales as vast as petabytes and zettabytes.
- Yet Another Resource Negotiator (YARN) – It oversees and monitors the nodes within the cluster and ensures efficient utilization of resources. Additionally, it is responsible for job and task scheduling.
- MapReduce – Developed using the Java programming language, it serves as a software framework for data processing. MapReduce integrates two distinct components. However, the presence of class imbalance can pose challenges.
- Map: Dividing a dataset into smaller portions, it enables the conversion of data into a different format, such as a key-value pair.
- Reduce: The transformation of key-value pairs into tuples takes place in the second section.
- Hadoop Common – It offers shared Java libraries that can be utilized across all modules.

4.2.2 Spark

The open-source data processing platform Spark was created with handling massive amounts of data in mind. Its main objectives are to give Big Data applications excellent computing speed, scalability, and programmability. Applications that use streaming data, graph data, machine learning, and artificial intelligence (AI) are given special consideration by Spark. The Spark analytics engine demonstrates data processing speeds that are 10 to 100 times faster compared to rival engines like Hadoop. It scales by distributing processing jobs over multiple mass computers that are developed with parallel and fault-tolerant capabilities. Data scientists and analysts commonly employ popular programming languages such as Scala, Java, Python, and R. Spark is regularly related to Apache Hadoop, particularly in relation to MapReduce, as it serves as a built-in data-processing function within Hadoop. Spark and MapReduce differ significantly in that Spark processes data and saves it in memory for later use without giving up data quality [13].

4.3 Jupyter

Users can create and share documents with live code, equations, graphics, and text using the free and open-source online tool known as the Jupyter Notebook. It provides a broad range of capabilities in areas including statistical modelling, data processing, machine learning, and data visualization. The Jupyter Notebook is a useful tool for a variety of applications thanks

to its flexible capabilities, which make activities like data extraction, transformation, and exploration easier. The Jupyter Notebook serves as a web-based interactive computational tool for composing Jupyter papers (previously known as IPython Notebooks). Depending on the context, the term "notebook" can have different meanings, the term "Jupyter" typically refers to the Jupyter online application, Jupyter Python web server, or the Jupyter document format. The main goal of Project Jupyter, as stated on its official website, is to offer open-source software, open standards, and services that enable interactive computing in various programming languages. Their aim is to provide a platform that supports collaborative and exploratory work across different domains [14].

5. Results and Analysis

5.1 Proposed System

In a business context, the prediction of customer attrition refers to the efforts made by a company to retain customers who have a high likelihood of discontinuing their use of its services. This study seeks to identify the clients who are most likely to abandon the product, website, or service. It's very difficult to maintain current clients because finding new ones might be difficult. Churn may be reduced by methodically looking back at the prior history of the key client. It is possible to identify new buyers who may churn by doing an effective analysis of the large amounts of customer data that are maintained. Drivers have a range of choices for predicting and preventing client churn by analysing the feasible data in a number of different ways. Figure 7 depicts the suggested system's steps. This number includes Data Collection, Data Pre-processing, Data Preparation, Data Visualisation Tools, and Data Prediction.

5.1.1 Collection of Data

The ideal data is selected from the standard telecom/network datasets, which can be utilized to create accurate prediction models.

5.1.2 Pre-processing of data

Data Management can be a challenging activity. Unprocessed data may result in multiple variations in the output. Data cleansing, data transformation, and the selection of features are all essential components of the pre-processing of data. Data transformation occurs when a pair of explanatory variables are in a convertible format (binomial form to binary form), which is advantageous for the provided model. Data Cleaning, as its name suggests, is the process of removing/treating absent or unnecessary data.

5.1.3 Preparation of data

As implied by its name, data preparation involves the process of refining data in a manner that enhances the analysis of the data. Each category is iterated until a desirable set is obtained. Prior to data examination, the given dataset must be cleansed to eliminate errors and redundancy. If these factors are not observed, it may lead to improper predictions or false values. Data preparation involves numerous steps like discretization of numerical values, imputing the absence of fields, feature selection of prominent use, transforming of one value set (discrete) into another, and deriving new variables. Imputation involves filling the gaps in the dataset, based on previous findings and removal of any discrepancy in the dataset. Deriving new variables occurs on the basis of discretization/transformation. The given study helps to predict the number of users/customers that churn from a particularly given service provider, and the outcomes are predicted in terms of probability predictions. The binary set of values i.e., 0 and 1 are utilized in this Logistic Regression Model.

5.1.4 Prediction of data

The company considers the finished product; hence it is very important to convey its output as presented in a pictorial/graphical format that is quick to grasp and assists the company in making the necessary predictions and analyses which lead to the prosperity of the company. A number of elements work together to achieve that goal.

5.1.5 Tools for Visualization of Data

The most effective approach of getting data across is to use data visualization tools. The patterns formed help to analyse the crucial essence of data, which is often ignored when one puts a sole statistical approach to study our data. Tableau and Google Charts are some great visualization tools.

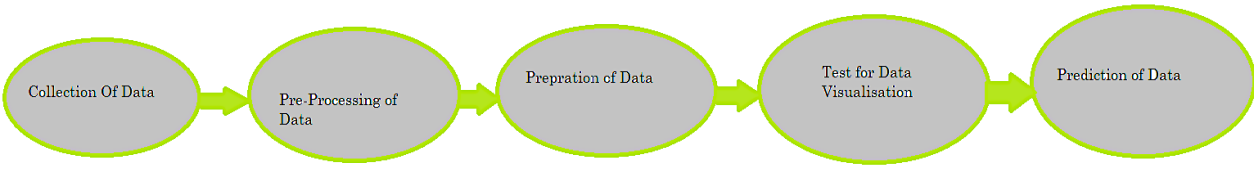


Figure 7. Steps for Proposed System

5.2 System Requirement

While working with large amounts of data, they want to build a platform on which the data can be processed and evaluated. The topic of conversation revolved around constructing a platform due to the limitations of a standard system, both in terms of capacity and processing power, to effectively manage substantial volumes of data. The data is as large as feasible, as measured by petabytes and gigabytes. As a result, two things are required.

- Operating System: Linux Programming Language
- Python

5.3 Vector Assembler

VectorAssembler is used to extract the turns of a sequence of columns into a single vector column. VectorAssembler takes various inputs such as boolean, numeric, etc. The contents of each input column are combined and arranged into a vector, following a specific desired order for each row. This vectorization process is employed in training machine learning models like logistic regression, allowing the integration of both raw features and also given features generated by various feature transformers into a single column. In addition to examining the dataset, ensuring data accuracy is crucial. Hence, after performing the necessary cleaning procedures, an evaluation process is conducted for each customer in the dataset, involving the assessment of missing values and data types. Within this dataset, one has the ability to make selections based on specific criteria or parameters ‘Age’, ‘Account Manager’, ‘Total Purchase’, ‘Num Sites’, and ‘Years’. Applying the feature selection algorithm Vector Assembler on the dataset had the following results [15] (Figures 8 and 9).

```
assembler = VectorAssembler(inputCols=[
 'Age',
 'Total_Purchase',
 'Account_Manager',
 'Years',
 'Num_Sites'
 ],
outputCol="features")
```

Figure 8. Feature selection

```
+--------------------+-----+
|            features|churn|
+--------------------+-----+
|[42.0,11066.8,0.0...|    1|
|[41.0,11916.22,0....|    1|
|[38.0,12884.75,0....|    1|
|[42.0,8010.76,0.0...|    1|
|[37.0,9191.58,0.0...|    1|
|[48.0,10356.02,0....|    1|
|[44.0,11331.58,1....|    1|
|[32.0,9885.12,1.0...|    1|
|[43.0,14062.6,1.0...|    1|
+--------------------+-----+
```

Figure 9. Feature/churn selection of Vector assembler

5.4 LogisticRegression

When dealing with decimal numbers as the desired data format, logistic regression emerges as the optimal predictive model to employ. Logistic regression serves as a prediction model that evaluates and describes the relationship within a dataset. In

the context of customer churn analysis, logistic regression was applied to ascertain the probability of churn based on specific client characteristics or variables outlined in our dataset. According to Hassouna et al. (2016), logistic regression is also employed to assess the likelihood of customer attrition. Logistic regression is a statistically focused tool for investigating the effect of variables on others. When they want to apply Logistic Regression, they want to train the desired dataset at the random split (0.7,0.3) and then apply logistic regression to find the prediction. They used the feature selection result of the vector assembler and found the raw prediction and probability and they got 70 percent at raw prediction and 99.30 percent at probability [16] (Figures 10 and 11).

features	churn	rawPrediction	probability	prediction
[22.0,11254.38,1....]	0.0	[4.96050524236784...]	[0.99303940400139...]	0.0
[27.0,8628.8,1.0,...]	0.0	[5.905974690516 68...]	[0.99728426899703...]	0.0
[28.0,8670.98,0.0...]	0.0	[8.19872388612018...]	[0.99972507132363...]	0.0
[28.0,9090.43,1.0...]	0.0	[1.70655875854888...]	[0.84638940648796...]	0.0
[28.0,11128.95,1....]	0.0	[4.43696894638154...]	[0.98830660639038...]	0.0
[28.0,11204.23,0....]	0.0	[1.81406789469421...]	[0.85985279779525...]	0.0
[28.0,11245.38,0....]	0.0	[3.69189769600253...]	[0.97568147324479...]	0.0
[29.0,5900.78,1.0...]	0.0	[4.59258813963566...]	[0.98997490494809...]	0.0
[29.0,8688.17,1.0...]	1.0	[2.96650688314298...]	[0.95103787662661...]	0.0
[29.0,9378.24,0.0...]	0.0	[4.92859535749842...]	[0.99281533194024...]	0.0
[29.0,10203.18,1....]	0.0	[4.01265512959848...]	[0.98223595571025...]	0.0
[29.0,11274.46,1....]	0.0	[4.78549371666795...]	[0.99171914494550...]	0.0
[29.0,13240.01,1....]	0.0	[7.03746026219900...]	[0.99912241630757...]	0.0
[29.0,13255.05,1....]	0.0	[4.30395241860729...]	[0.98666518428846...]	0.0
[30.0,6744.87,0.0...]	0.0	[3.71004655959537...]	[0.97610839642025...]	0.0
[30.0,7960.64,1.0...]	1.0	[3.57503617850392...]	[0.97274900827954...]	0.0
[30.0,8403.78,1.0...]	0.0	[6.44804766880066...]	[0.99841889297270...]	0.0
[30.0,8677.28,1.0...]	0.0	[4.41655478245109...]	[0.98806831982257...]	0.0
[30.0,8874.83,0.0...]	0.0	[3.23928343658757...]	[0.96228611290728...]	0.0
[30.0,10183.98,1....]	0.0	[3.10250401931621...]	[0.95699591531282...]	0.0

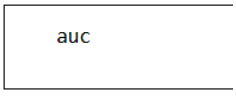
Figure 10. Prediction and Raw prediction of Logistic regression

summary	churn	prediction
count	624	624
mean	0.18269230769230768	0.14102564102564102
stddev	0.3867240627102176	0.34832721924783666
min	0.0	0.0
max	1.0	1.0

Figure 11. Churn/Prediction of Logistic Regression

5.5 BinaryClassificationEvaluator

Binary classification is a machine learning technique within the realm of supervised learning, aimed at categorizing incoming observations into distinct classes. Numerous applications can be found where binary classification is employed, wherein each observation is assigned to one of the two classes represented by the 0 or 1 columns. The BinaryClassificationEvaluator algorithm is also used to find the accuracy and prediction with the dataset they used earlier in this project. They find the accuracy and prediction of the most accurate algorithms and compare them. After the comparison and applying Classification, they got 75.69 percent accuracy and a probability of around 99.21 percent [17] (Figures 12 and 13). In the realm of modelling," True Positive" denotes the scenario where the best model accurately predicts and classifies a situation as positive (TP). On the other hand," True Negative" is the term utilized when the model correctly predicts and identifies outcomes as negative (TN). In a similar vein, a false positive arises when an individual without the condition is mistakenly diagnosed as positive, resulting in a positive test result (FP).



0.7569444444444445

Figure 12. Accuracy of BinaryClassificationEvaluator

features	churn	rawPrediction	probability	prediction
[25.0,9672.03,0.0...	0	[4.83695271757974...	[0.99213122290547...	0.0
[26.0,8787.39,1.0...	1	[0.81425607830479...	[0.69301570648089...	0.0
[26.0,8939.61,0.0...	0	[6.70259561188525...	[0.99877378437479...	0.0
[29.0,9617.59,0.0...	0	[4.55236239781693...	[0.98956771010833...	0.0
[29.0,12711.15,0....	0	[5.36391356043500...	[0.99533927925365...	0.0

Figure 13. Prediction and Rawprediction of BinaryClassificationEvaluator

5.6 AUC-ROC Curve

AUC, which stands for" Area Under Curve" of the ROC (Receiver Operating Characteristic), is a metric used to assess the performance of Binary Classification models. The ROC curve showcases the relation between the True Positive Rate (TPR) and False Positive Rate (FPR). It provides insights into the model's ability to discriminate between different classes. A higher AUC value indicates superior performance in terms of distinction, with values closer to 1 being desirable. The Y-axis represents the true positive rate, and the X-axis represents the false positive rate. The AUC curve, depicted in Figure 14 corresponds to a model accuracy of 75.69 percent.

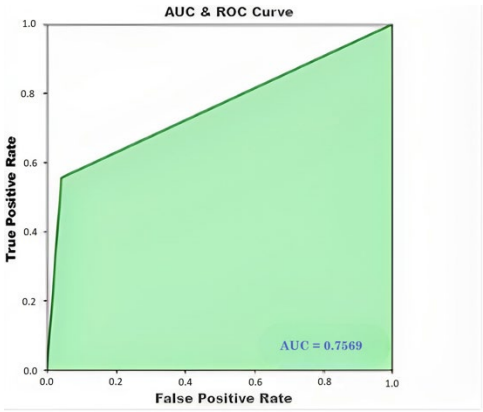


Figure 14. AUC-ROC

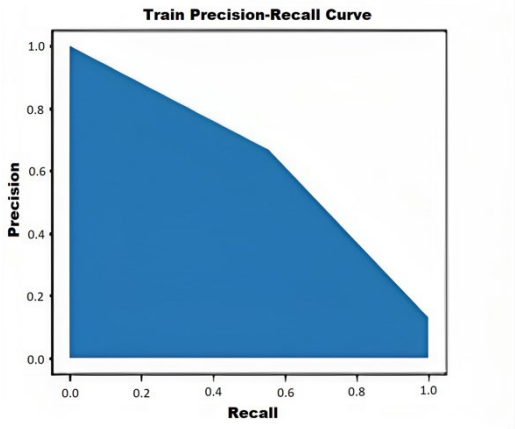


Figure 15. Precision-Recall Curve

5.7 Precision-Recall Curve

Just like the AUC-ROC curve, Precision-Recall is another evaluator for Binary Classification. It represents the classifier's performance about numerous thresholds graphically in contrast to the f1-score. The Y-axis plots the Precision value, while the X-axis plots the Recall value (Figure 15).

5.8 MulticlassClassificationEvaluator

MulticlassClassificationEvaluator was also used in which they find the test error, predictions, feature measure, threshold measure, and area under ROC. In this classification they find the highest accuracy of 91 percent and the test error and area under ROC are 0.94 percent. [18] (Figures 16 and 17).

```
Test Error = 0.0942029
```

Figure 16. Test error of MulticlassClassificationEvaluator

```
+--------------------+--------------------+
|                 FPR|                 TPR|
+--------------------+--------------------+
|                 0.0|                 0.0|
|0.00196078431372549|0.043859649122807015|
|0.00392156862745098| 0.08771929824561403|
|0.00392156862745098| 0.14035087719298245|
|0.00392156862745098| 0.19298245614035087|
+--------------------+--------------------+
only showing top 5 rows

areaUnderROC: 0.9185758513931891
```

Figure 17. FPR AND TPR and area Under ROC of MulticlassClassificationEvaluator

5.9 Classification report/Confusion matrix

The use of a confusion matrix is a widely adopted method for evaluating Logistic Regression models. This matrix is structured as an N x N matrix, where N represents the number of classes. It provides a breakdown of correct and incorrect predictions made by the Regression model. The confusion matrix primarily works on four values:

The classification report is another performance parameter used in this research to evaluate the performance of the model. It comprises of mainly three parameters:

- Precision: It produces the result of how many positive predictions made are correct.
- Recall: It quantifies the accuracy of positive case predictions among all positive cases, as depicted.
- F1-Score: Illustrates how it provides the weighted average or harmonic mean of recall and precision. The classification report and confusion matrix will find the precision, recall, F1-score, accuracy, array, and confusion matrix with normalization / without normalization [10] (Figures 18-21).
- Normalization: Normalization in the term of Confusion matrix means a max of 1.00 samples is being taken for each given group. Thus, the row sum concludes to be 100.

	precision	recall	f1-score	support
0	0.93	0.96	0.95	240
1	0.67	0.56	0.61	36
accuracy			0.91	276
macro avg	0.80	0.76	0.78	276
weighted avg	0.90	0.91	0.90	276

Figure 18. Precision/ Recall/ F1-Score and accuracy

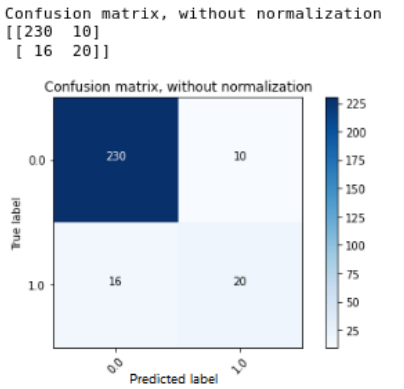


Figure 19. Confusion matrix without Normalisation

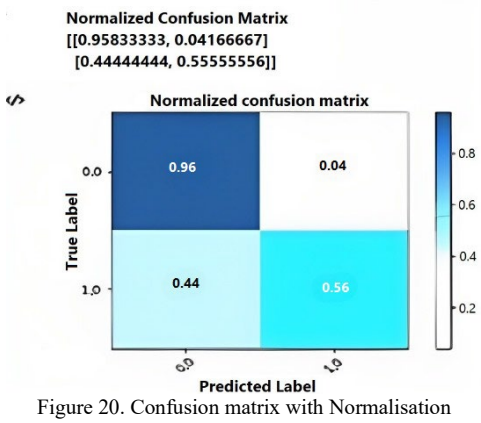


Figure 20. Confusion matrix with Normalisation

```
array([[230,  10],
       [ 16,  20]])
```

Figure 21. The array form of data in a confusion matrix

6. Conclusion

The objective of telecommunication sector research is to assist businesses in increasing their profitability. In the telecom industry, churn prediction plays a pivotal role in revenue generation as it enables the anticipation of customer attrition. Accurate prediction models with high AUC values are crucial in this context. To assess and train the model, the sample data is segregated into 20 percent for testing and 80 percent for training. A 5-fold cross-validation approach is employed for evaluating performance and adjusting hyperparameters. The preparation of data for machine learning algorithms involves the utilization of a classifier, evaluators, efficient feature transformation, and a selection process. They also observed that the data was not matched. Customer attrition represents only 5 percent of the data. Massive reductions or the usage of tree methodologies not affected by this vulnerability were utilized to solve the issue. The well-known algorithms Logistic Regression, Vector Assembler, BinaryclassificationEvaluator, and MulticlassificationEvaluator algorithms were selected for their variety and scope for this type of prediction. Because the AUC value was 91.301 percent, the technique of preparation, feature selection, and characteristic integration had the most influence on the performance of this model. In each and every measurement, the classification report, and confusion matrix produced the best results. The AUC percentage was 91.21 percent. In terms of AUC values, the classification report, and confusion matrix comes in second, followed by the Random Forest and Logistic regression. When evaluating models using a new dataset from different time periods, the MulticlassEvaluator yielded the highest accuracy of 91 percent. The phenomenon of non-stationary data models is particularly relevant to customer churn in telecom companies, as it emphasizes the importance of training on fresh data for each instance. Additionally, incorporating social network analysis features enhances the predictive outcomes for telecom churn prediction.

References

[1] Huang, Y., Zhu, F., Yuan, M., Deng, K., Li, Y., Ni, B., Dai, W., Yang, Q., and Zeng, J. Telco Churn Prediction with Big Data. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, pp. 607-618, 2015.

[2] Burez, J. and Van den Poel, D. Handling Class Imbalance in Customer Churn Prediction. *Expert Systems with Applications*, vol. 36, no. 3, pp. 4626-4636, 2009.

[3] Lalwani, P., Mishra, M.K., Chadha, J.S., and Sethi, P. Customer Churn Prediction System: A Machine Learning Approach. *Computing*, pp. 1-24, 2022.

[4] Hadden, J., Tiwari, A., Roy, R., and Ruta, D. Computer Assisted Customer Churn Management: State-of-The-Art and Future Trends. *Computers & Operations Research*, vol. 34, no. 10, pp. 2902-2917, 2007.

[5] Kisioglu, P. and Topcu, Y.I. Applying Bayesian Belief Network Approach to Customer Churn Analysis: A Case Study on the Telecom Industry of Turkey. *Expert Systems with Applications*, vol. 38, no. 6, pp. 7151-7157, 2011.

[6] Coussement, K. and Van den Poel, D. Churn Prediction in Subscription Services: An Application of Support Vector Machines While Comparing Two Parameter-Selection Techniques. *Expert systems with applications*, vol. 34, no. 1, pp. 313-327, 2008.

[7] International Centre for Mechanical Sciences; International Federation for the Theory of Machines and Mechanisms, Nevins, J.L., and Whitney, D.E. *The force vector assembler concept*, Springer Berlin Heidelberg, pp. 273-288, 1972.

[8] Allison, P.D. *Logistic regression using SAS: Theory and application*. SAS institute, 2012.

[9] Kakarla, R., Krishnan, S., and Alla, S. Model Evaluation. In *Applied Data Science Using PySpark*, Springer, pp. 205–249, 2021.

[10] Marom, N.D., Rokach, L., and Shmilovici, A. using the Confusion Matrix for Improving Ensemble Classifiers. In *2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel*, IEEE, pp. 000555-000559, 2010.

[11] Ahmad, A.K., Jafar, A., and Aljoumaa, K. Customer Churn Prediction in Telecom using Machine Learning in Big Data Platform. *Journal of Big Data*, vol. 6, no. 1, pp. 1-24, 2019.

[12] Prajapati, V. *Big data analytics with R and Hadoop*. Packt Publishing Ltd, 2013.

[13] Dong, G., Fu, X., Li, H., and Pan, X. An Accurate Sequence Assembly Algorithm for Livestock, Plants and Microorganism based on Spark. *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 31, no. 08, pp. 1750024, 2017.

[14] Lovrić, M., Molero, J.M., and Kern, R. PySpark and RDKit: Moving Towards Big Data in Cheminformatics. *Molecular informatics*, vol. 38, no. 6, pp. 1800082, 2019.

[15] Khan, M.A., Karim, M.R., and Kim, Y. A Two-Stage Big Data Analytics Framework with Real World Applications using Spark Machine Learning and Long Short-Term Memory Network. *Symmetry*, vol. 10, no. 10, pp. 485, 2018.

[16] Chaudhuri, K. and Monteleoni, C. Privacy-Preserving Logistic Regression. *Advances in neural information processing systems*, vol. 21, 2008.

[17] Erraissi, A. and Banane, M. Machine Learning Model to Predict the Number of Cases Contaminated by COVID-19. International Journal of Computing and Digital Systems, vol. 9, pp. 1–11, 2020.

[18] Branitskiy, A., Kotenko, I., and Saenko, I. Applying Machine Learning and Parallel Data Processing for Attack Detection in IoT. *IEEE Transactions on Emerging Topics in Computing*, vol. 9, no. 4, pp. 1642-1653, 2020.