

**A SIMPLIFIED TACTIC PREDICTION OF CUSTOMER CHURN FOR
TELECOMMUNICATION**

at

**Sathyabama Institute of Science and Technology
(DEEMED TO BE UNIVERSITY)**

Submitted in partial fulfillment of the requirements for the
award of Bachelor of Engineering Degree in

Computer Science and Engineering By

CHERUKUMALLI DIVYA

(Reg. No. 37110163)

DIKHITA KASHYAP

(Reg No. 37110191)



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING



SCHOOL OF COMPUTING
SATHYABAMA INSTITUTE OF SCIENCE AND
TECHNOLOGY JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI - 600119, TAMILNADU
APRIL-2021

SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY

(ESTABLISHED UNDER SECTION 3 OF UGC ACT,1956) JEPPIAAR NAGAR, RAJIV GANDHI
SALAI, CHENNAI - 600119 www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide
work of **CHERUKUMALLI DIVYA (Reg. No. 37110163) DIKHITA KASHYAP**
(Reg. No. 37110191) who carried out the project entitled "**A SIMPLIFIED**
TACTIC PREDICTION OF CUSTOMER CHURN FOR
TELECOMMUNICATION" under my supervision from April 2021 to June 2021

Internal Guide

Mrs.Jithina Jose M.E(Ph.D).,

Head of the Department

Dr. S. VIGNESHWARI M.E., Ph.D.,

Dr.Lakshman L M.E., Ph.D.,

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I CHERUKUMALLI DIVYA (REGNO: 37110163) DIKHITA KASHYAP (REGNO: 37110191) hereby declare that the professional Training Report entitled **“A SIMPLIFIED TACTIC PREDICTION OF CUSTOMER CHURN FOR TELECOMMUNICATION”** done by me under the guidance of **Mrs.Jithina Jose M.E,(Ph.D).**, at Sathyabama Institute of Science and technology is submitted in partial fulfillment of the requirements of the award of Bachelor of Engineering degree in Computer Science and Engineering

DATE:

**PLACE:
SIGNATURE OF THE CANDIDATE**

ACKNOWLEDGEMENT

We are pleased to acknowledge our sincere thanks to Board of management of SATHYABAMA for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

We convey your thanks to **Dr. T. SASIKALA M.E., Ph.D., Dean of the Department, Department of Computer Science and Engineering** for providing us the necessary support and details at the right time during the progressive reviews.

We convey our thanks to **Dr. S. VIGNESHWARI M.E., Ph.D., and Dr.Lakshman L M.E,Ph.D., Head of the Department, Department of Computer Science and Engineering** for providing us the necessary support and details at the right time during the progressive reviews.

We would like to express our sincere and deep sense of gratitude to our Project Guide **Mrs Jithina Jose M.E(Ph.D).**, for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

We wish to express our thanks to all Teaching and Non-teaching staff members of the Department of **COMPUTER SCIENCE AND ENGINEERING** who were helpful in many ways for the completion of the project.



TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	v
	LIST OF FIGURES	
	INTRODUCTION	1
	1.1 Overview	1
	1.2 Scope of Project	1
	1.3 Domain Overview	1
	1.4 Deep Learning	2
2	LITERATURE SURVEY	11
	1.1 Literature Survey 1	11
	1.1 Literature Survey 2	14
	1.2 Literature Survey 3	16
2	SYSTEM ANALYSIS	18
	2.1 Existing System	18
	2.2 Drawbacks of Existing Systems	19
	2.3 Proposed System	19
	2.4 Advantages of Proposed System	20

4	REQUIREMENT SPECIFICATION	21
4.1	Introduction	21
4.2	Hardware and Software Specification	21
4.2.1	Hardware Specification	21
4.2.2	Software Specification	21
4.3	Software Description	22
4.3.1	About Python	22
4.3.2	About IDLE	32
5	SYSTEM DESIGN	36
5.1	Introduction	36
5.2	Architecture Design	36
6	SYSTEM IMPLEMENTATIONS	38
6.1	List of Modules	38
6.1.1	Module 1	38
6.1.2	Module 2	39
6.1.3	Module 3	40
6.1.4	Module 4	41


```

iiiiiiiiiiiiiiiiiiiiiiikl111111

```




for a specific variable. This means that a data frame's rows do not need to contain, but can contain, the same type of values: they can be numeric, character, logical, etc.

Now, Data Frames in Python are very similar: they come with the Pandas library, and they are defined as two-dimensional labeled data structures with columns of potentially different types.

In general, you could say that the Pandas Data Frame consists of three main components: the data, the index, and the columns.

Firstly, the Data Frame can contain data that is:

- a Pandas Data Frame
- a Pandas Series: a one-dimensional labeled array capable of holding any data type with axis labels or index. An example of a Series object is one column from a Data Frame.

NumPy ndarray, which can be a record or structured

- a two-dimensional ndarray
- dictionaries of one-dimensional ndarray's, lists, dictionaries or Series. Some of the key features of Python Pandas are as follows: It provides Data Frame objects with default and customized indexing which is very fast and efficient.
- There are tools available for loading data of different file formats into in-memory data objects
 - It is easy to perform data alignment and integrated handling of missing data in Python Pandas.
 - It is very simple to perform pivoting and reshaping of data sets in Pandas.
 - It also provides indexing, label-based slicing, and sub-setting of large data sets.
 - We can easily insert and delete columns from a data structure.
 - Data aggregation and transformations can be done using group by.
 - High-performance merging and joining of data can be done using Pandas.
 - It also provides time series functionality.

- Inserting and deleting columns in data structures.
- Merging and joining data sets.
- Reshaping and pivoting data sets.
- Aligning data and dealing with missing data.
- Manipulating data using integrated indexing for Data Frame objects.
- Performing split-apply-combine on data sets using the group by engine.
- Manipulating high-dimensional data in a data structure with a lower dimension using hierarchical axis indexing.
- Sub setting, fancy indexing, and label-based slicing data sets that are large in size.
- Generating data range, converting frequency, date shifting, lagging, and other time-series functionality.
- Reading from files with CSV, XLSX, TXT, among other formats.
- Arranging data in an order ascending or descending.
- Filtering data around a condition. • Analyzing time series.
- Iterating over a data set.

NUMPY

NumPy is the core library for scientific computing in Python. It provides a high-performance multidimensional array object, and tools for working with these arrays. If you are already familiar with MATLAB, you might find this tutorial useful to get started with NumPy.

A NumPy array is a grid of values, all of the same type, and is indexed by a tuple of nonnegative integers. The number of dimensions is the rank of the array; the shape of an array is a tuple of integers giving the size of the array along each dimension.

NumPy is, just like SciPy, Scikit-Learn, Pandas, etc. one of the packages that you just can't miss when you're learning data science, mainly because this library provides you

with an array data structure that holds some benefits over Python lists, such as: being more compact, faster access in reading and writing items, being more convenient and more efficient.

NumPy is a Python library that is the core library for scientific computing in Python. It contains a collection of tools and techniques that can be used to solve on a computer mathematical models of problems in Science and Engineering. One of these tools is a high-performance multidimensional array object that is a powerful data structure for efficient computation of arrays and matrices. To work with these arrays, there's a vast amount of high-level mathematical functions operate on these matrices and arrays.

an array is basically nothing but pointers. It's a combination of a memory address, a data type, a shape, and strides:

- The data pointer indicates the memory address of the first byte in the array,
- The data type or dtype pointer describes the kind of elements that are contained within the array,
- The shape indicates the shape of the array, and
- The strides are the number of bytes that should be skipped in memory to go to

the next element. If your strides are (10,1), you need to proceed one byte to get to the next column and 10 bytes to locate the next row.

Or, in other words, an contains information about the raw data, how to locate an element and how to interpret an element.

With NumPy, we work with multidimensional arrays. We'll dive into all of the possible types of multidimensional arrays later on, but for now, we'll focus on 2-dimensional arrays. A 2- dimensional array is also known as a matrix, and is something you should be familiar with. In fact, it's just a different way of thinking about a list of lists. A matrix has rows and columns. By specifying a row number and a column number, we're able to extract an element from a matrix.

towards GNUplot is the fact that Matplotlib is a Python module. Due to the growing interest in python the popularity of matplotlib is continually rising as well.

Another reason for the attractiveness of Matplotlib lies in the fact that it is widely considered to be a perfect alternative to MATLAB, if it is used in combination with NumPy and Scipy. Whereas MATLAB is expensive and closed source, Matplotlib is free and open source code. It is also object-oriented and can be used in an object oriented way. Furthermore it can be used with general-purpose GUI toolkits like wxPython, Qt, and GTK+. There is also a procedural "pylab", which designed to closely resemble that of MATLAB. This can make it extremely easy for MATLAB users to migrate to matplotlib.

Matplotlib can be used to create publication quality figures in a variety of hardcopy formats and interactive environments across platforms.

Another characteristic of matplotlib is its steep learning curve, which means that users usually make rapid progress after having started. The official website has to say the following about this: "matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatterplots, etc., with just a few lines of code."

Deep Learning

Deep learning is a computer software that mimics the network of neurons in a brain. It is a subset of machine learning and is called deep learning because it makes use of deep neural networks. Deep learning algorithms are constructed with connected layers.

- The first layer is called the Input Layer
- The last layer is called the Output Layer
- All layers in between are called Hidden Layers. The word deep means the network join neurons in more than two layers.

A deep neural network provides state-of-the-art accuracy in many tasks, from object detection to speech recognition.

To grasp the idea of deep learning, imagine a family, with an infant and parents. The toddler points objects with his little finger and always says the word 'cat.' As its parents are concerned about his education, they keep telling him 'Yes, that is a cat' or 'No, that is not a cat.' The infant persists in pointing objects but becomes more accurate with 'cats.' The little kid, deep down, does not know why he can say it is a cat or not. He has just learned how to hierarchies complex features coming up with a cat by looking at the pet overall and continue to focus on details such as the tails or the nose before to make up his mind.

The learning occurs in two phases.

- The neural network repeats these two phases hundreds to thousands of time until it has reached a tolerable level of accuracy. The repeat of this two-phase is called an iteration
- ### Classification of Neural Networks

- 9

2. Deep neural network: Deep neural networks have more than one layer. For instance, Google LeNet model for image recognition counts 22 layers.

The computational models in Deep Learning are loosely inspired by the human brain. The multiple layers of training are called Artificial Neural Networks (ANN).

Neuron

Artificial Neural Networks contain layers of neurons. A neuron is a computational unit that calculates a piece of information based on weighted input parameters. Inputs accepted by the neuron are separately weighted.

Inputs are summed and passed through a non-linear function to produce output. Each layer of neurons detects some additional information, such as edges of things in a picture or tumors in a human body. Multiple layers of neurons can be used to detect additional information about input parameters.

Nodes

Artificial Neural Network is an interconnected group of nodes akin to the vast network of layers of neurons in a brain. Each circular node represents an artificial neuron and an arrow represents a connection from the output of one neuron to the input of another.

Inputs

Inputs are passed into the first layer. Individual neurons receive the inputs, with each of them receiving a specific value. After this, an output is produced based on these values.

Outputs

The outputs from the first layer are then passed into the second layer to be processed. This continues until the final output is produced. The assumption is that the correct output is predefined.

Each time data is passed through the network, the end result is compared with the correct one, and tweaks are made to their values until the network creates the correct final output each time.

Some of the commonly used neural networks are as follows:

1. Artificial Neural Network (ANN)
2. Convolutional Neural Network (CNN)
3. Recurrent Neural Network (RNN)
4. Deep Neural Network (DNN)
5. Deep Belief Network (DBN)

Artificial neural networks are one of the main tools used in machine learning. As the “neural” part of their name suggests, they are brain-inspired systems which are intended to replicate the way that we humans learn. Neural networks consist of input and output layers, as well as (in most cases) a hidden layer consisting of units that transform the input into something that the output layer can use. They are excellent tools for finding patterns which are far too complex or numerous for a human programmer to extract and teach the machine to recognize.

Generally, the working of a human brain by making the right connections is the idea behind ANNs. That was limited to use of silicon and wires as living neurons and dendrites.

Here, neurons, part of human brain. That was composed of 86 billion nerve cells. Also, connected to other thousands of cells by Axons. Although, there are various inputs from sensory organs. That was accepted by dendrites.

As a result, it creates electric impulses. That is used to travel through the Artificial neural network. Thus, to handle the different issues, neuron send a message to another neuron. As a result, we can say that ANNs are composed of multiple nodes. That imitate biological neurons of the human brain. Although, we connect these neurons by links. Also, they interact with each other. Although, nodes are used to take input data. Further, perform simple operations on the data.

LITERATURE SURVEY 2

Title

Comparing oversampling techniques to handle the class imbalance problem: A

customer churn prediction
case study.

Authors

Adnan Amin, Sajid Anwar, Awais Adnan,
Muhammad Nawaz,
Authors Newton Howard, Junaid Qadir, Ahmad
Hawalah, Amir Hussain

2016

Published year

✓ High reported accuracy and interpretability for understanding key
drivers, as well as for providing information to set
up retention actions

Efficiency

✓ The weight also provides important.
information, specifically, outliers.
✓ Especially useful for data expressing categorized values

Drawbacks

↓ It is difficult to identify risk factors which can be
addressed by the retention process.
↓ Does not separate the training data based on the hardness of fitting a
base learner.
↓ Does not provides what the customer really needs

In this study, we survey six well-known sampling
techniques and compare the performances of these key techniques
i.e., Mega-trend Diffusion Function (MTDF), Synthetic Minority
Oversampling Technique (SMOTE), Adaptive Synthetic Sampling
approach (ADASYN), Couples Top-N Reverse k-Nearest Neighbor
(TRkNN), Majority Weighted Minority Oversampling Technique
(MWMOTE) and Immune centroids oversampling technique
(ICOTE). Moreover, this study also reveals the evaluation of four

Title	A Big Data Clustering Algorithm for Mitigating the Risk of Customer Churn
Author	Wenjie Bi, Meili Cai, Mengqi Liu and Guo Li
	Published Year 2016
✓ Attempts to “boost” the accuracy of any given learning algorithm	
✓ Trained with reweighted training data	
✓ Separating customers based on the assigned weight	
Drawbacks	↓ The excess variables not only burden the computational process in generation of a model ↓ Does not build a distinct prediction model for each defined cluster.
↓ Does not provides an opportunity to define a high risk customer group	
This paper proposes	a new clustering algorithm called Semantic Driven Subtractive Clustering Method (SDSCM) is proposed. Experimental results indicate that SDSCM has stronger clustering semantic strength than Subtractive Clustering Method (SCM) and fuzzy c-means (FCM). Then a parallel SDSCM algorithm is implemented through a Hadoop MapReduce framework. In the case study, the proposed parallel SDSCM algorithm enjoys a fast running speed when compared with the other methods. Furthermore, we provide some marketing strategies in accordance with the clustering results, and a simplified marketing activity is simulated to ensure profit maximization. Axiomatic Fuzzy Set (AFS) AFS is an effective way to describe the fuzzy concept. The membership

19

The existing system profit maximizing framework to solve this special customer churn problem. Specifically, we first build a profit maximizing classifier to predict whether a customer is prone to churn or not, in which profitable customers have more probability to be classified correctly. Then, we describe and quantify the switching behaviors of 4G customers by introducing a notion of switching score. Finally, we investigate relation between operators and customers, finding that some features of 4G service plans provided by operators indeed affect switching behaviors of 4G customers remarkably. This provides insight into reasonable design of 4G service plans for operators in the future. Our framework uncovers the root causes of intraoperative customer churn and solve the problem well.

The existing system propose the concept of changing trend to reflect changing of switching score of each customer over several months. Through aggregating changing trend of customers using the same 4G service plan, we can obtain design evaluation score of this 4G service plan. Then, we can analyze relations between switching behaviors of 4G customers and 4G service plans by design evaluation score.

3.2 Drawbacks of Existing System

- ↓ Not an effective model for forecasting customer churn for mid-sized and small sized companies
- ↓ Irrelevant variables are not eliminated.
- ↓ Cannot predict a risk level and timeline will be a more challenging issue.
- ↓ Heavyweight
- ↓ Big payloads

3.3 Proposed System

CHAPTER 4

4.1 INTRODUCTION

Clearly defined requirements are essential signs on the road that leads to a successful project. They establish a formal agreement between a client and a provider that they are both working to reach the same goal. High-quality, detailed requirements also help mitigate financial risks and keep the project on a schedule. According to the Business Analysis Body of Knowledge definition, requirements are a usable representation of a need. Creating requirements is a complex task as it includes a set of processes such as elicitation, analysis, specification, validation, and management. In this article, we'll discuss the main types of requirements for software products and provide a number of recommendations for their use.

4.2 Hardware & Software Specification

4.2.1 Hardware Specification

Hardware Requirement

Hardware	Minimum Requirement
Disk Space	32 GB or more,10 GB or more for Foundation Edition
Processor	1.4 GHz 64 bit
Memory	512 MB
Display	(800 × 600) Capable video adapter and monitor

4.2.2 Software Specification

Python coding style comprises physical lines as well as logical lines or statements. A physical line in a Python program is a sequence of characters, and the end of the line terminates the line sequence as opposed to some other languages, such as C and C++ where a semi-colon is used to mark the end of the statement. A logical line, on the other hand, is composed of one or more physical lines. The use of a semi-colon is not prohibited in Python, although it's not mandatory. The NEWLINE token denotes the end of the logical line. A logical line that only contains spaces, comments, or tabs are called blank lines and they are ignored by the interpreter.

As we saw that in Python, a new line simply means that a new statement has started. Although, Python does provide a way to split a statement into a multiline statement or to join multiple statements into one logical line. This can be helpful to increase the readability of the statement. Following are the two ways to split a line into two or more lines:

Explicit Line Joining

In explicit line joining, we use a backward slash to split a statement into a multiline statement.

Implicit Line Joining

Statements that reside inside [], {}, or () parentheses can be broken down into two or more physical lines without using a back slash.

Multiple Statements on a Single Line

In Python, it is possible to club multiple statements in the same line using a semi-colon; however, most programmers do not consider this to be a good practice as it reduces the readability of the code.

Whitespaces and Indentation

Unlike most of the programming languages, Python uses indentation to mark a block of code. According to Python coding style guideline or PEP8, we should keep an indent size of four.

Pandas is a popular Python package for data science, and with good reason: it offers powerful, expressive and flexible data structures that make data manipulation and analysis easy, among many other things. The Data Frame is one of these structures. Pandas is a high-level data manipulation tool developed by Wes McKinney. It is built on the NumPy package and its key data structure is called the Data Frame. Data Frames allow you to store and manipulate tabular data in rows of observations and columns of variables.

Pandas is built on top of the NumPy package, meaning a lot of the structure of NumPy is used or replicated in Pandas. Data in pandas is often used to feed statistical analysis in SciPy, plotting functions from Matplotlib, and machine learning algorithms in Scikit-learn.

Jupyter Notebooks offer a good environment for using pandas to do data exploration and modeling, but pandas can also be used in text editors just as easily.

Jupyter Notebooks give us the ability to execute code in a particular cell as opposed to running the entire file. This saves a lot of time when working with large datasets and complex transformations. Notebooks also provide an easy way to visualize pandas' Data Frames and plots. As a matter of fact, this article was created entirely in a Jupyter Notebook.

There are two types of data structures in pandas:

Series and Data Frames.

1. Series: a pandas Series is a one dimensional data structure ("a one dimensional ndarray") that can store values — and for every value it holds a unique index, too.
2. Data Frame: a pandas Data Frame is a two (or more) dimensional data structure — basically a table with rows and columns. The columns have names and the rows have indexes.

Those who are familiar with R know the data frame as a way to store data in rectangular grids that can easily be overviewed. Each row of these grids corresponds to

- Data aggregation and transformations can be done using group by.
- High-performance merging and joining of data can be done using Pandas.
- It also provides time series functionality.
- Inserting and deleting columns in data structures.
- Merging and joining data sets.
- Reshaping and pivoting data sets.
- Aligning data and dealing with missing data.
- Manipulating data using integrated indexing for Data Frame objects.
- Performing split-apply-combine on data sets using the group by engine.
- Manipulating high-dimensional data in a data structure with a lower dimension using hierarchical axis indexing.
- Sub setting, fancy indexing, and label-based slicing data sets that are large in size.
 - Generating data range, converting frequency, date shifting, lagging, and other time-series functionality.
- Reading from files with CSV, XLSX, TXT, among other formats.
- Arranging data in an order ascending or descending.
- Filtering data around a condition.
- Analyzing time series.
- Iterating over a data set.

NumPy

NumPy is the core library for scientific computing in Python. It provides a high-performance multidimensional array object, and tools for working with these arrays. If you are already familiar with MATLAB, you might find this tutorial useful to get started with NumPy.

A NumPy array is a grid of values, all of the same type, and is indexed by a tuple of nonnegative integers. The number of dimensions is the rank of the array; the shape of an array is a tuple of integers giving the size of the array along each dimension.

NumPy is, just like SciPy, Scikit-Learn, Pandas, etc. one of the packages that you just can't miss when you're learning data science, mainly because this library provides you with an array data structure that holds some benefits over Python lists, such as: being more compact, faster access in reading and writing items, being more convenient and more efficient.

NumPy is a Python library that is the core library for scientific computing in Python. It contains a collection of tools and techniques that can be used to solve on a computer mathematical models of problems in Science and Engineering. One of these tools is a high-performance multidimensional array object that is a powerful data structure for efficient computation of arrays and matrices. To work with these arrays, there's a vast amount of high-level mathematical functions operate on these matrices and arrays.

an array is basically nothing but pointers. It's a combination of a memory address, a data type, a shape, and strides:

- The data pointer indicates the memory address of the first byte in the array,
- The data type or dtype pointer describes the kind of elements that are contained within the array,
- The shape indicates the shape of the array, and
- The strides are the number of bytes that should be skipped in memory to go to the next element. If your strides are (10,1), you need to proceed one byte to get to the next column and 10 bytes to locate the next row.

Or, in other words, an array contains information about the raw data, how to locate an element and how to interpret an element.

With NumPy, we work with multidimensional arrays. We'll dive into all of the possible types of multidimensional arrays later on, but for now, we'll focus on 2-dimensional

lead developer was Michael Droettboom. One of the free and open-source Python libraries which is basically used for technical and scientific computing is Python SciPy. Matplotlib is widely used in SciPy as most scientific calculations require plotting of graphs and diagrams.

Matplotlib is a plotting library like GNUplot. The main advantage towards GNUplot is the fact that Matplotlib is a Python module. Due to the growing interest in python the popularity of matplotlib is continually rising as well.

Another reason for the attractiveness of Matplotlib lies in the fact that it is widely considered to be a perfect alternative to MATLAB, if it is used in combination with Numpy and Scipy. Whereas MATLAB is expensive and closed source, Matplotlib is free and open source code. It is also object-oriented and can be used in an object oriented way. Furthermore it can be used with general-purpose GUI toolkits like wxPython, Qt, and GTK+. There is also a procedural "pylab", which designed to closely resemble that of MATLAB. This can make it extremely easy for MATLAB users to migrate to matplotlib.

Matplotlib can be used to create publication quality figures in a variety of hardcopy formats and interactive environments across platforms.

Another characteristic of matplotlib is its steep learning curve, which means that users usually make rapid progress after having started. The official website has to say the following about this: "matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatterplots, etc., with just a few lines of code.

4.3 About IDE

IDE- Integrated Development Environment. It is software suite. It contains different instruments for web application software engineers to compose and test applications.

Programming Web application software engineers utilize wide scope of devices habitually all through application source code creation, building and application testing.

PyCharm is an Integrated Development Environment (IDE) that provides tools for writing efficient code. It supports various frameworks such as Django, Flask, and Pyramid. PyCharm is majorly used by developers for application development with Python language. Let's suppose, we are making a project using Python programming language.

We know that Python is extensively popular among the developer's community, as well as in industries, due to its following features:

- Simple to code: Python is a very simple language compared to other programming languages. Also, it is a developer-friendly language.
- Object-oriented language: Python is an object-oriented language that uses concepts of classes and objects.
- GUI programming support: We can create graphical user interfaces using Python Tkinter library. Besides, we can use modules such as PyQt5, PyQt4, and wxPython to create a GUI application.
- High-level language: Python is a high-level programming language. We do not need to take care of managing memory or the system architecture while writing Python codes.
- Portability: Python is a portable programming language, i.e., we can run Python codes on any platform. For example, we can run a code written for Windows in Linux, UNIX, and Mac, without changing it.
- Integrated language: Python is an integrated programming language. We can easily integrate C, C++, etc. with Python.
- Interpreted language: Python is an interpreted programming language as well. Python's source code is changed to bytecode. Due to this, it is directly executed without compiling.
- Large set of standard libraries: Python consists of a rich set of modules, packages, and libraries. It helps in various stages of application development. "notebook" or "notebook documents" denote documents that contain both code and rich text elements, such as figures, links, equations, ... Because of the mix

of code and text elements, these documents are the ideal place to bring together an analysis description, and its results, as well as, they can be executed perform the data analysis in real time.

The Jupyter Notebook App produces these documents.

As a server-client application, the Jupyter Notebook App allows you to edit and run your notebooks via a web browser. The application can be executed on a PC without Internet access, or it can be installed on a remote server, where you can access it through the Internet. Its two main components are the kernels and a dashboard.

A kernel is a program that runs and introspects the user's code. The Jupyter Notebook App has a kernel for Python code, but there are also kernels available for other programming languages. The dashboard of the application not only shows you the notebook documents that you have made and can reopen but can also be used to manage the kernels: you can which ones are running and shut them down if necessary.

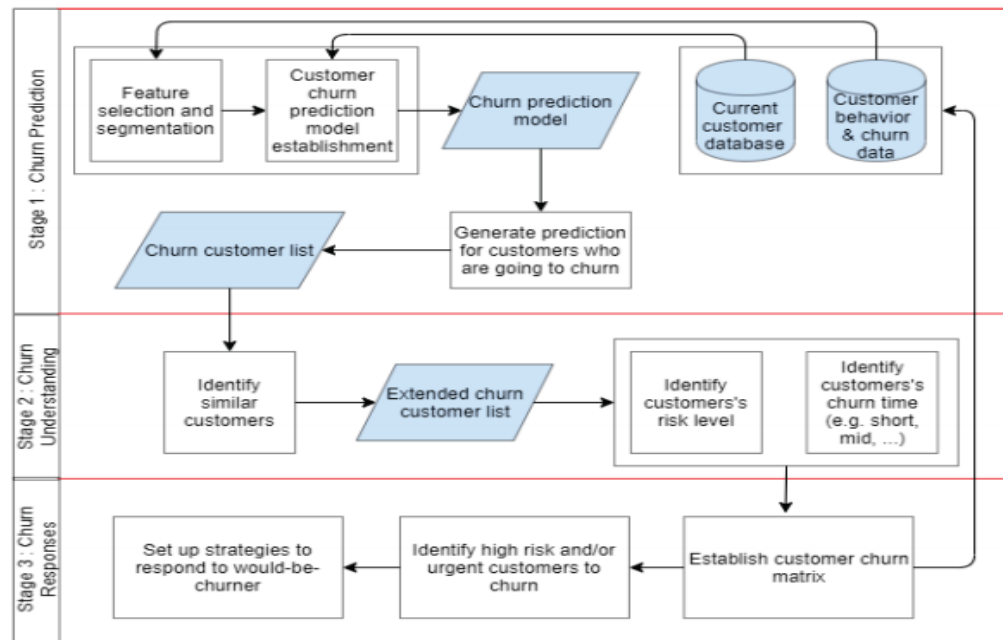


Fig. 5.1 Architecture

1.1. Module 1: Customer Churn Data Analysis

Ordinal features – Features which has some order. For example, t-shirt size feature can have values in ['small', 'medium', 'large', 'extra-large']. You may note that there is an order to the values.

Nominal features – Features which are just labels or names and don't have any order. For example, if the color of car is a feature, color can take value such as ['white', 'red', 'black', 'blue']. You may note there is no order to the value of the color features.

Scaling of Features is an essential step in modeling the algorithms with the datasets. The data that is usually used for the purpose of modeling is derived through various means. Standardization: Standard Scaler standardizes a feature by subtracting the mean and then scaling to unit variance. Unit variance means dividing all the values by the standard deviation. Standardization can be helpful in cases where the data follows a Gaussian distribution (or Normal distribution). However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

6.1.2 Module 2: Customer Churn Model Creation and Prediction

The procedure involves taking a dataset and dividing it into two subsets. The first subset is used to fit the model and is referred to as the training dataset. The second subset is not used to train the model; instead, the input element of the dataset is provided to the model, then

6.1.1. Module 1: Customer Churn Data Analysis

6.1.2 Module 2: Customer Churn Model Creation and Prediction

Project Information 41 © WISEN IT SOLUTIONS

predictions are made and compared to the expected values. This second dataset is referred to as the test dataset.

Train Dataset: Used to fit the machine learning model.

Test Dataset: Used to evaluate the fit machine learning model.

The objective is to estimate the performance of the machine learning model on new data: data not used to train the model. By default, the program ignores the original order of data. It randomly picks data to form the training

tree-like model of decisions. A decision tree is drawn upside down with its root at the top. In the image on the left, the bold text in black represents a condition/internal node, based on which the tree splits into branches/ edges. The end of the branch that doesn't split anymore is the decision/leaf, in this case, whether the passenger died or survived, represented as red and green text respectively.

Although, a real dataset will have a lot more features and this will just be a branch in a much bigger tree, but you can't ignore the simplicity of this algorithm. The feature importance is clear and relations can be viewed easily. This methodology is more commonly known as learning decision tree from data and above tree is called Classification tree as the target is to classify passenger as survived or died. Regression trees are represented in the same manner, just they predict continuous values like price of a house. In general, Decision Tree algorithms are referred to as CART or Classification and Regression Trees.

If the testing or development team gets software that is not working correctly and developed by someone else, so to find the error, the team will first need a document. Now, if the documents are available then the team will quickly find out the cause of the error by examining documentation. But, if the documents are not available then the tester need to do black box and white box testing again, which will waste the time and money of the organization. More than that, Lack of documentation becomes a problem for acceptance.

46

key features don't work or the key bugs have not been fixed yet. Here our focus will be on primary and core application work flow.

To conduct smoke testing, we don't write test cases. We just pick the necessary test cases from already written test cases.

Do we really write test cases for all testing types? Here in this article, we have given clear idea on choosing testing types to write test cases.

As mentioned earlier, here in Smoke Testing, our main focus will be on core application work flow. So we pick the test cases from our test suite which cover major functionality of the application. In general, we pick minimal number of test cases that won't take more than half an hour to execute.

After receiving a Software build with the minor issues fixes in code or functionality, Sanity testing is carry out to check whether the bugs reported in previous build are fixed & there is regression introduced due to these fixes i.e. not breaking any previously working functionality.

The main aim of Sanity testing to check the planned functionality is working as expected. Instead of doing whole regression testing the Sanity testing is perform.

Sanity tests helps to avoid wasting time and cost involved in testing if the build is failed. Tester should reject the build upon build failure.

After completion of regression testing the Sanity testing is started to check the defect fixes & changes done in the software application is not breaking the core functionality of the software. Typically this is done nearing end of SDLC i.e. while releasing the software. You can say that sanity testing is a subset of acceptance testing. We can also say Tester Acceptance Testing for Sanity testing.

7.8 Regression Testing

Regression Testing is a type of testing that is done to verify that a code change in the software does not impact the existing functionality of the product. This is to make sure the product works fine with new functionality, bug fixes or any change in the existing

feature. Previously executed test cases are re-executed in order to verify the impact of change.

Regression Testing is a Software Testing type in which test cases are re-executed in order to check whether the previous functionality of the application is working fine and the new changes have not introduced any new bugs.

This test can be performed on a new build when there is a significant change in the original functionality that too even in a single bug fix.

For regression testing to be effective, it needs to be seen as one part of a comprehensive testing methodology that is cost-effective and efficient while still incorporating enough variety—such as well-designed frontend UI automated tests alongside targeted unit testing, based on smart risk prioritization—to prevent any aspects of your software applications from going unchecked. These days, many Agile work environments employing workflow practices such as XP (Extreme Programming), RUP (Rational Unified Process), or Scrum appreciate regression testing as an essential aspect of a dynamic, iterative development and deployment schedule.

But no matter what software development and quality-assurance process your organization uses, if you take the time to put in enough careful planning up front, crafting a clear and diverse testing strategy with automated regression testing at its core, you can help prevent projects from going over budget, keep your team on track, and, most importantly, prevent unexpected bugs from damaging your products and your company's bottom line.

Performance testing is the practice of evaluating how a system performs in terms of responsiveness and stability under a particular workload. Performance tests are typically executed to examine speed, robustness, reliability, and application size. The process incorporates “performance” indicators such as:

- Browser, page, and network response times
- Server request processing times

- Georges D. Olle Olle and Shuqin Cai. A hybrid churn prediction model in mobile telecommunication industry. A Hybrid Churn Prediction Model in Mobile Telecommunication Industry, (1):55–62, 2014.
- S.H.Dolatabadi and F. Keynia, “Designing of customer and employee churn prediction model based on data mining method and neural predictor,” in 2017 2nd International Conference on Computer and Communication Systems (ICCCS), July 2017, pp. 74–77.
- P.Praveen, B.Rama ,Uma N.Dulhare,” A study on monothetic Divisive Hierarchical Clustering Method”, International Journal of Advanced Scientific Technologies ,Engineering and Management Sciences (IJASTEMS-ISSN: 2454-356X) Volume.3,Special Issue.1,March.2017
- Wenjie Bi , Meili Cai , Mengqi Liu “A Big Data Clustering Algorithm for Mitigating the Risk of Customer Churn” IEEE Transactions on Industrial Informatics , Volume: 12, Issue: 3, June 2016 Amjad Hudaib, Reham Dannoun, Osama Harfoushi, Ruba Obiedat, and Hossam Faris. Hybrid data mining models for predicting customer churn. Network and System Sciences, 8:91–96, 5 2015