Project Report

On

Predicting Customer Conversion Rate on Bank Telemarketing Data using Machine Learning

Submitted in partial fulfilment for the award of

Post Graduate Diploma in Big Data Analysis (PG-DBDA)

From Know-IT (Pune)

Guided By:

Mr. Tushar Kute

Mrs. Trupti Joshi

Submitted By:

Chetan Amrao (220943025004)

Mandar Patil (220943025017)

Kaustubh Patil (220943025025)

Prince Ganer (220943025028)

CERTIFICATE

TO WHOMSOEVER IT MAY CONCERN

This is to certify that

Chetan Amrao (220943025004)

Mandar Patil (220943025017)

Kaustubh Patil (220943025025)

Prince Ganer (220943025028)

Have successfully completed their project on

Predicting Customer Conversion Rate on Bank Telemarketing Data using Machine Learning

Under guidance of Mr. Tushar Kute & Mrs. Trupti Joshi

ACKNOWLEDGEMENT

This project **"Predicting Customer Conversion Rate on Bank Telemarketing Data using Machine Learning"** was a great learning experience for us and we are submitting this work to CDAC Know-IT (Pune).

We all are very glad to mention the name of our guide of **Mr. Tushar Kute & Mrs. Trupti Joshi** for their valuable inputs and suggestions throughout this project. Their experience and support helped us to overcome various obstacles and challenges during the course of project work.

We are highly grateful to **Mr. Amit Patil** (training coordinator at Know-IT) for his support and patience during the course of 6 months of Post Graduate Diploma in Big Data Analytics (PG-DBDA) through C-DAC ACTS, Pune.

Our most heartful thanks goes to **Mrs. Bakul Joshi** (course coordinator PG-DBDA) who gave all the technical help and support when we struggled on basics during the course. Her participation really pushed us to finish the project with more hope.

From:

Chetan Amrao (220943025004)

Mandar Patil (220943025017)

Kaustubh Patil (220943025025)

Prince Ganer (220943025028)

TABLE OF CONTENTS

1. Abstract
2. Introduction
3. System Requirements
    a. Software requirements
    b. Hardware requirements
4. Functional Requirements
5. System Architecture
6. Data Pre-Processing
7. Machine Learning Algorithms
8. Data Visualization and Representation
9. Conclusion and Future Scope

ABSTRACT

The project aims to develop a predictive model for customer conversion rates in a telemarketing campaign using machine learning algorithms. We employed data mining and prescriptive analysis on a large dataset of over 1.8 million data points from a Portuguese bank telemarketing campaign. The dataset is downloaded using an API and loaded into a NoSQL database using ETL processes. Cloud technology is used for in-memory computing and stream processing. Using machine learning algorithms in the Spark package, we explored the impact of various factors such as holidays and customer deposit preferences on customer conversion rates. The aim is to identify critical questions about the campaign's success. To facilitate decision-making, we will develop a Flask-based web framework and GUI for users to interact with the data and make informed decisions. The project provides valuable insights to support data-driven decisions, which can improve the efficiency of future telemarketing campaigns.

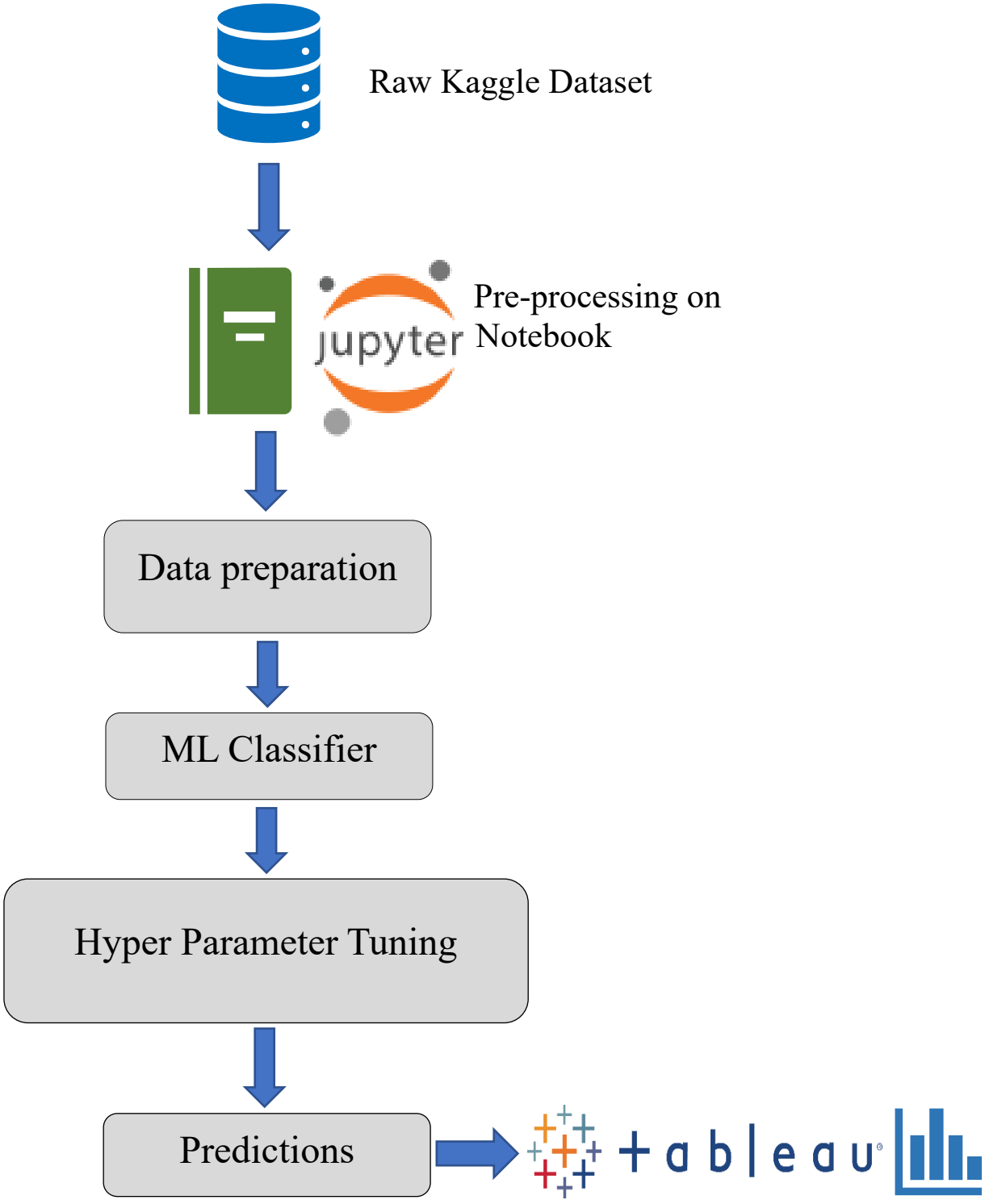
INTRODUCTION

Telemarketing is a technique used by marketing professionals to attract new customers and keep hold of old customers. It is used in sectors like banking, financial services and insurance (BFSI), information technology and telecommunications, and consulting to locate potential consumers through phone calls where they are asked to purchase goods or services.

According to Pecan AI's 2022 market research, 84% of marketing leaders use predictive analysis and 40% of marketing professionals a lack of knowledge and training regarding content marketing as another struggle and almost 42% of financial marketers say that hiring top talent is a problem they face in their organization.

We propose two solutions based on telemarketing bank dataset. A model for customer segmentation based on prediction for the part of the month deposit was successfully made and a web tool for marketers to decide which customers to prioritize.

SYSTEM ARCHITECTURE



SYSTEM REQUIREMENTS

Software requirements:

1. Arrow 1.2.3
2. Flask 2.2.3
3. Git 2.38.windows.1
4. Jupyter Notebook 6.4.12
5. Kaggle 1.5.13
6. MongoDB 6.0.4
7. PyMongo 4.3.3
8. PySpark 3.3.2
9. Python 3.9
10. Tableau 2022.04

Minimum hardware requirements:

1. 8GB RAM
2. Intel/AMD 4 core
3. Storage up to 1GB
4. Network connection for downloading dataset, python libraries
5. Platform:
    a. Windows 10 (or)
    b. Ubuntu 20.04

FUNCTIONAL REQUIREMENTS

1. Arrow
    a. *version 1.2.3*
    b. Built on the Apache Arrow columnar memory format, which enables fast, efficient, and interoperable access to data across different programming languages.
    c. Provides support for distributed processing using popular frameworks like Apache Spark and Dask.

2. Flask
    a. *version 2.2.3*
    b. Lightweight and flexible web framework for building web applications using Python.
    c. Provides the essential tools for building web applications with Python, including routing, templating, and request handling.

3. Git
    a. version 2.38.windows.1
    b. Git is a popular version control system that is used to manage code changes, collaborate with others, and track project history.
    c. Git allows developers to create multiple branches of a codebase, which can be used for testing, development, or experimentation.
    d. It allows developers to rollback changes to their codebase, either to a previous version or to a specific point in time.

4. Jupyter Notebook
    a. *version 6.4.12*
    b. Used for data exploration because of its ability to display data visualizations inline with code.
    c. Supports collaborative coding where multiple users can work on the same notebook simultaneously.
    d. Jupyter Notebook supports many programming languages, including Python, R, and Julia.

5. Kaggle
    a. *version 1.5.13*
    b. Kaggle API is a command-line tool that allows us to download and upload data, scripts.
    c. Before using the Kaggle API, we need to authenticate our account by creating an API token on the Kaggle website.
    d. You can also use the Kaggle API to upload datasets to Kaggle. This is useful if you have created a new dataset that you want to share with the Kaggle community.

6. MongoDB
    a. *version 6.0.4*
    b. MongoDB is designed to scale horizontally, meaning that you can add more servers to your cluster to handle increasing amounts of data and traffic.
    c. MongoDB's query language is powerful and expressive, allowing you to perform complex queries and aggregation operations on your data. It also supports geospatial queries, text search, and graph processing.
    d. There are also many third-party tools and libraries that integrate with MongoDB.

7. PyMongo
    a. *version 4.3.3*
    b. PyMongo is a Python driver for MongoDB, which allows Python developers to interact with MongoDB databases and collections using a simple and intuitive API.
    c. PyMongo allows you to perform Create, Read, Update, and Delete operations on MongoDB databases and collections using a simple and intuitive syntax.
    d. PyMongo allows you to create and manage indexes on your MongoDB collections, which can improve query performance and make your applications faster and more efficient.

8. PySpark
    a. *version 3.3.2*
    b. PySpark is the Python API for Apache Spark, a popular big data processing framework. PySpark provides an easy-to-use interface for Python developers to perform distributed data processing using Spark.
    c. Allows developers to write distributed data processing code in Python.
    d. PySpark is a powerful tool for distributed data processing and is widely used in the big data community.

9. Python
    a. version 3.9
    b. General-purpose high-level language with great support for machine learning libraries.
    c. Well-suited for building complex applications, as well as for rapid prototyping and experimentation.

10. Tableau
    a. version 2022.04
    b. Tableau is a data visualization and business intelligence software that is used to analyse and present data in an intuitive and interactive way.
    c. Tableau's interactive dashboards allow users to explore data in real-time and create custom views and filters to analyse specific data subsets.

DATA PRE-PROCESSING

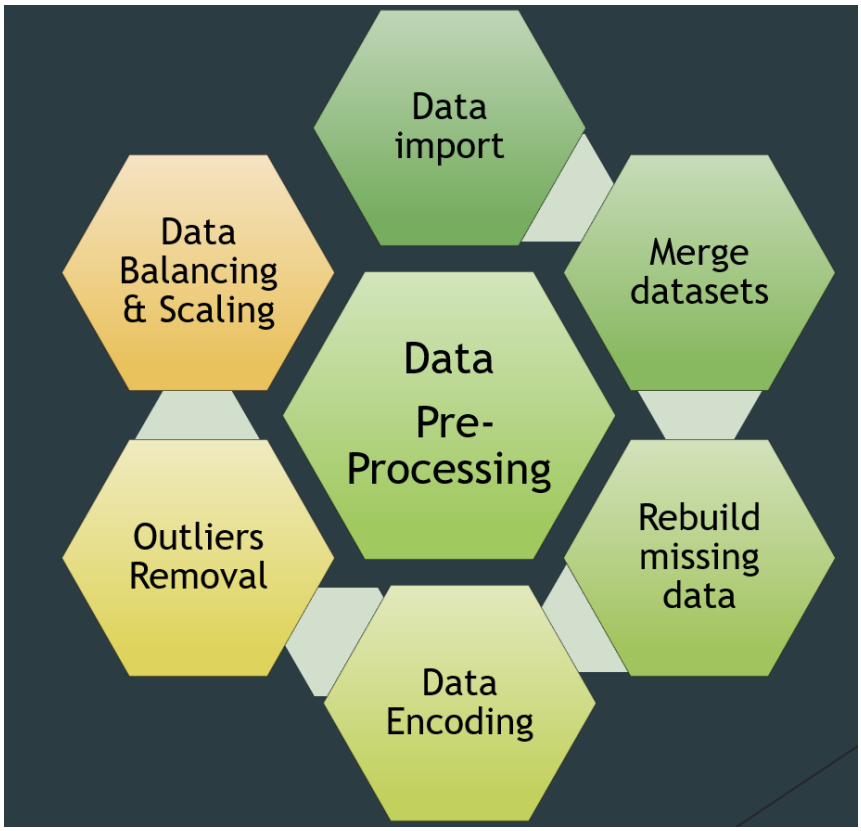


Fig. Data Cleaning Process

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting. After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleaning differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data.

MACHINE LEARNING ALGORITHMS

Machine learning is a subfield of artificial intelligence (AI) that focuses on the development of algorithms and models that enable computers to learn from and make predictions or decisions based on data, without being explicitly programmed to do so. In other words, it involves teaching machines to learn from experience, similar to the way humans learn.

The core idea behind machine learning is to create models or algorithms that can identify patterns in data, learn from these patterns, and use this knowledge to make predictions or decisions about new data. This is done by feeding the machine learning model with a large amount of training data, and then using various techniques to optimize the model's parameters so that it can accurately make predictions or decisions on new data.

In Predicting Customer Conversion Rate on Bank Telemarketing Data using Machine Learning project we applied various types of Classification Algorithms such as Logistic Regression, Support Vector Classifier, Decision Tree and Ensemble Algorithm of Random Forest. After the implementation we analysed the accuracy of all the applied algorithms.

We have used Supervised Learning Classification Algorithms for our project. These machine learning algorithms are explained as follows:

1. Logistic Regression:

Logistic regression is a popular statistical method used for binary classification problems, where the goal is to predict the probability of an event occurring based on input variables. It is a type of generalized linear model that uses a logistic function to model the relationship between the input variables and the output variable.

In logistic regression, the output variable is binary, meaning it can take only two possible values (e.g. 0 or 1, true or false, yes or no). The logistic function, also known as the sigmoid function, maps the input variables to the output probability. The logistic function has an S-shaped curve, which allows it to model nonlinear relationships between the input variables and the output probability.

2. Support Vector Machines:

Support Vector Classifier (SVC) is a machine learning algorithm used for binary and multi-class classification problems. It is a type of supervised learning algorithm that works by finding the optimal hyperplane that separates the different classes in the feature space.

In SVC, the algorithm finds a hyperplane that maximizes the margin between the two classes. The margin is defined as the distance between the hyperplane and the nearest data points from each class. The hyperplane that maximizes the margin is known as the optimal hyperplane, and the data points closest to the hyperplane are called support vectors.

The SVC algorithm has several advantages, including:

1. It can handle non-linearly separable data by transforming the input data into a higher-dimensional space.
2. It is effective in high-dimensional spaces, where the number of features is greater than the number of samples.
3. It is robust to outliers, since the algorithm only considers the support vectors, which are the data points closest to the hyperplane.

3. Decision Tree Classifier

Decision Tree Classifier is a machine learning algorithm used for classification problems. It is a type of supervised learning algorithm that works by recursively splitting the data into subsets based on the value of one of the input features. The goal of the algorithm is to create a decision tree that can accurately classify new data based on the input features.

The decision tree is created by finding the feature that best splits the data into two subsets, such that the resulting subsets are as homogeneous as possible with respect to the target variable. The process is repeated for each subset until a stopping criterion is met, such as a maximum depth or a minimum number of samples required for a leaf node.

The decision tree classifier can handle both categorical and numerical input features, and it can also handle missing data by using surrogate splits. The classifier can be used for both binary and multi-class classification problems.

The decision tree classifier has several advantages, including:

1. It is easy to interpret and visualize the decision tree, making it useful for understanding the relationship between the input features and the output variable.
2. It can handle both categorical and numerical input features.
3. It can handle missing data by using surrogate splits.
4. It is computationally efficient and can handle large datasets.

4. Random Forest Classifier

Random Forest Classifier is a popular ensemble machine learning algorithm used for classification problems. It is an extension of the decision tree algorithm, where multiple decision trees are trained on different subsets of the data, and the output is the majority vote of the individual trees.

The random forest algorithm works by creating multiple decision trees on different random subsets of the training data, using a technique called bootstrap aggregating or bagging. Each tree is trained on a different subset of the features, selected randomly at each node. The output of the random forest classifier is the majority vote of the individual trees.

Random Forest Classifier has several advantages over the decision tree algorithm, including:

1. It is less prone to overfitting compared to the decision tree algorithm, since it uses multiple trees trained on different subsets of the data.
2. It can handle both categorical and numerical input features.
3. It is computationally efficient and can handle large datasets.
4. It can provide feature importance measures, which can be useful for feature selection.

Random Forest Classifier is widely used in various fields such as image classification, bioinformatics, and finance. However, it may require careful tuning of the hyperparameters to achieve good performance, and the resulting model may be difficult to interpret compared to the decision tree algorithm.

DATA VISUALIZATION AND REPRESENTATION

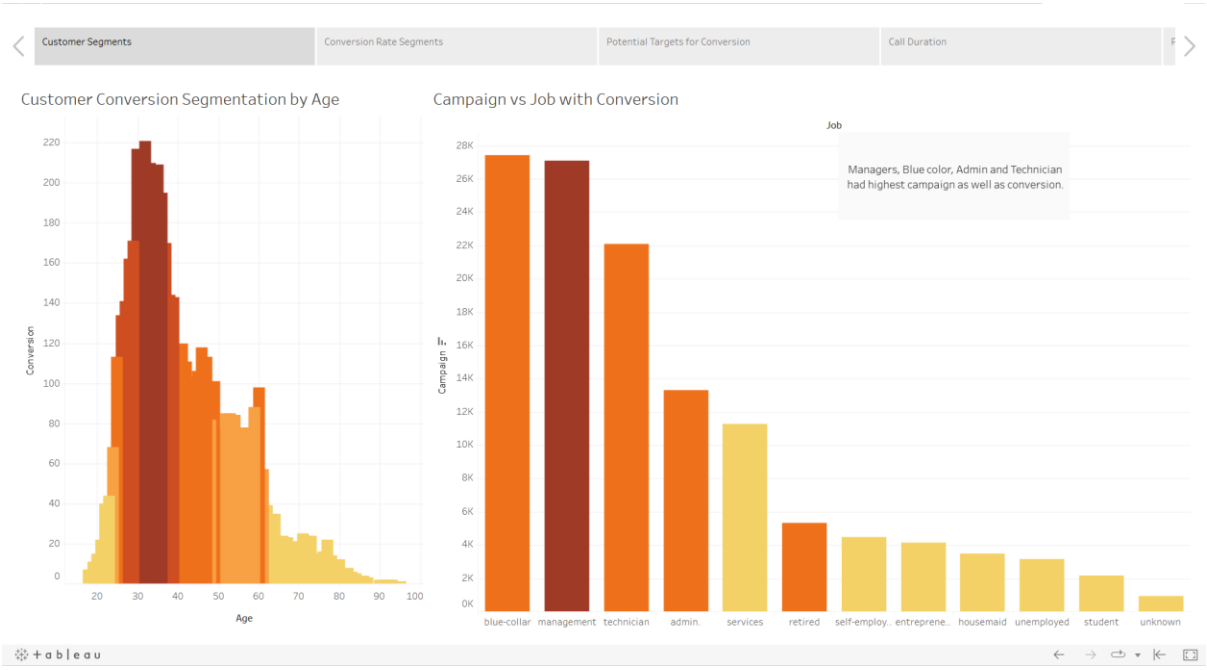


Fig. Customer Conversion Segmentation by Age and Job

It was analysed from these above visualisations that customers having job designations of blue-collar, technician and management jobs and having age between 25 to 60 years are of prime interest for term deposit subscription.

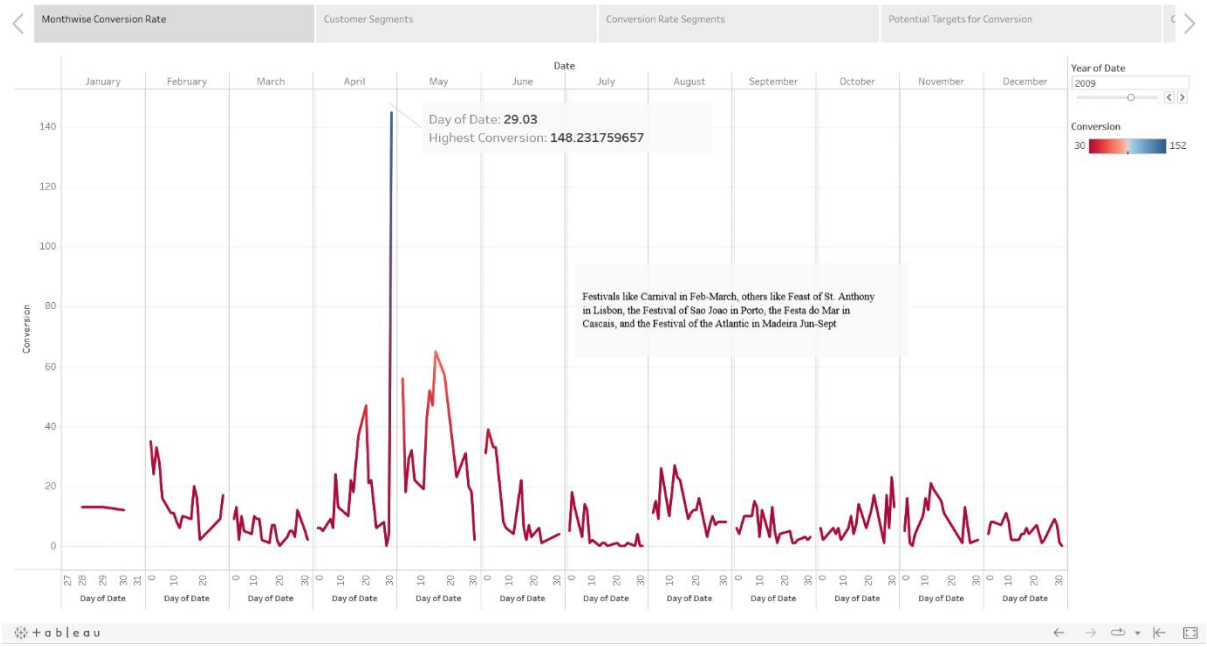


Fig. Month wise Conversion Rate

Above visualisations explain a clear trend in month wise distribution of the subscription for long-term deposits. Customers mostly brought subscriptions in mid-month interval.

For Quarter-wise trend, Long-term deposit subscriptions were brought mostly in the Quarter 2 of the fiscal year.

A decline in subscriptions was observed with respect to Quarter 2 where public festive holidays occur.

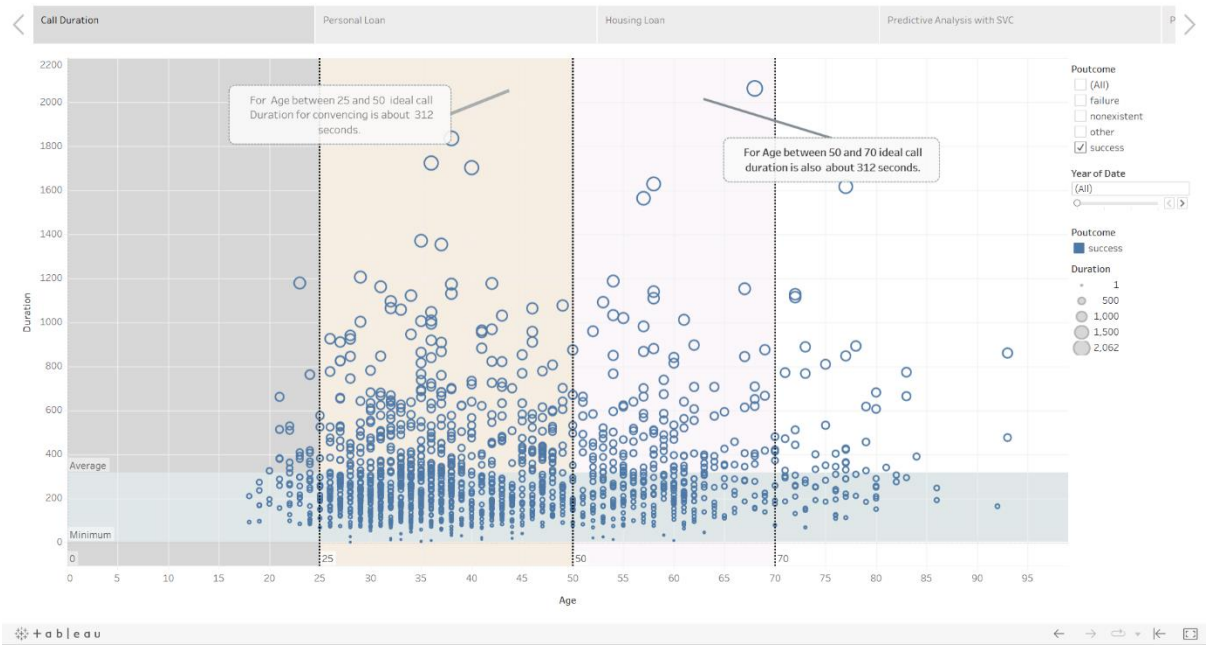


Fig. Age vs Call Duration

Most of the customers who were contacted and who purchased long-term deposit subscriptions had a call duration of about 320 seconds whereas customers who declined to purchase the long-term deposits had a call duration of 240 seconds. Marketing strategies (like lucrative offers) can be designed to boost the conversion rate.

CONCLUSION AND FUTURE SCOPE

In Predicting Customer Conversion Rate on Bank Telemarketing Data using Machine Learning, we combined two different datasets of raw data and pre-processed them with ETL to generate a single training set for various ML algorithms.

The results comparing the models can be seen in table below:

Model	**Accuracy**
Support Vector Classifier	0.85
Logistic Regression	0.86
Decision Tree Classifier	0.87
Random Forest Classifier	0.84
Random Forest Classifier with hyper parameter tuning	0.89

The ensembling model obtained with highest accuracy of 89% can be useful for management training tools and produce suggestions based on input by the marketing professionals. Here lies the future scope of this project, deploying the ML model in a distributed cloud computing infrastructure such as AWS, Azure or GCP with Flask server integrated with Hive to continuously churn predictions and suggestions for bank telemarketing conversion rate.