

**Car Buying Prediction Based on Available Data**

Submitted in Partial Fulfilment of the Requirement for the

APSIT Skills Internship under the

**Course: Machine Learning**

**Computer Engineering**

#### by

**Miss.Hardika Lalwani…. (20202003)**

**Mr. Jainam Zaveri……. (20202007)**

**Mr. Jayesh Jain………... (19102021)**

Under the Guidance of

**Prof. Jaya Gupta**

Department of Computer Engineering

Academic Year: July-2021





**CERTIFICATE**

#### This to certify that,

**Miss.Hardika Lalwani…. (20202003)**

**Mr. Jainam Zaveri……. (20202007)**

**Mr. Jayesh Jain……….. (19102021)**

#### have satisfactorily carried out the Project work entitled “Car Buying Prediction Based on Available Data” in partial fulfilment of APSIT Skills internship in Computer Engineering as laid down by University of Mumbai during the academic year 2021-2022.

**Prof. Adesh Hardas Prof. A. M. Deshpande**

Faculty guide Head of Department

**Contents**

[Introduction 1](#_TOC_250006)

[Hardware and Software 2](#_TOC_250005)

[Datasets and Solution 3](#_TOC_250004)

Output………………………………………………………………..4

Applications………………………………………………………….5

Advantages of Logistic Regression… 6

[Disadvantages of Logistic Regression 7](#_TOC_250001)

Conclusion/References………………………………………………8

# Introduction

Assume a firm is about to launch a new campaign for their new vehicle brand and wants to know which demographics are most likely to buy their new automobile so that they can create ads that specifically target those demographics. They called a social network advertising company for help, as they had data from a previous successful campaign. They now aim to create a model that will assist them in achieving their goal.

###### Project Aim:

The aim of the project is to Use Logistic Regression model to predict if a person is going to buy a new car or not based on the available data

1

# Hardware and Software

**Chrome:** Google Chrome is a cross-platform web browser developed by Google.

**Google Colab:** Colab is the commonly used abbreviation of the New York City artists' group Collaborative Projects, which was formed after a series of open meetings between artists of various disciplines.

**NumPy:** NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays

**Matplotlib:** Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy.

**Pandas:** pandas is a software library written for the Python programming language.

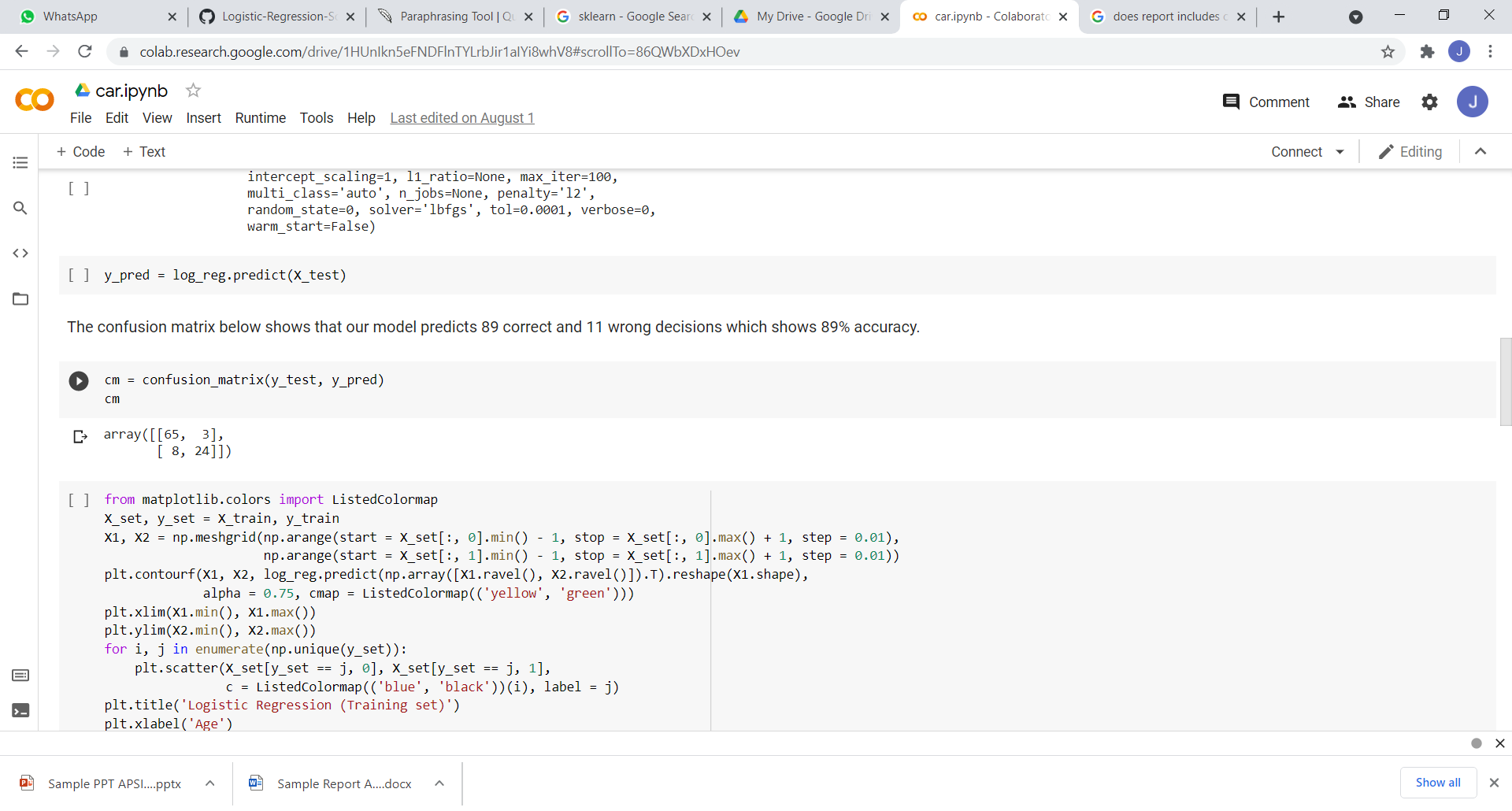
**Sklearn:** The sklearn library contains a lot of efficient tools for machine learning and statistical modeling.

2

# Datasets and Solution

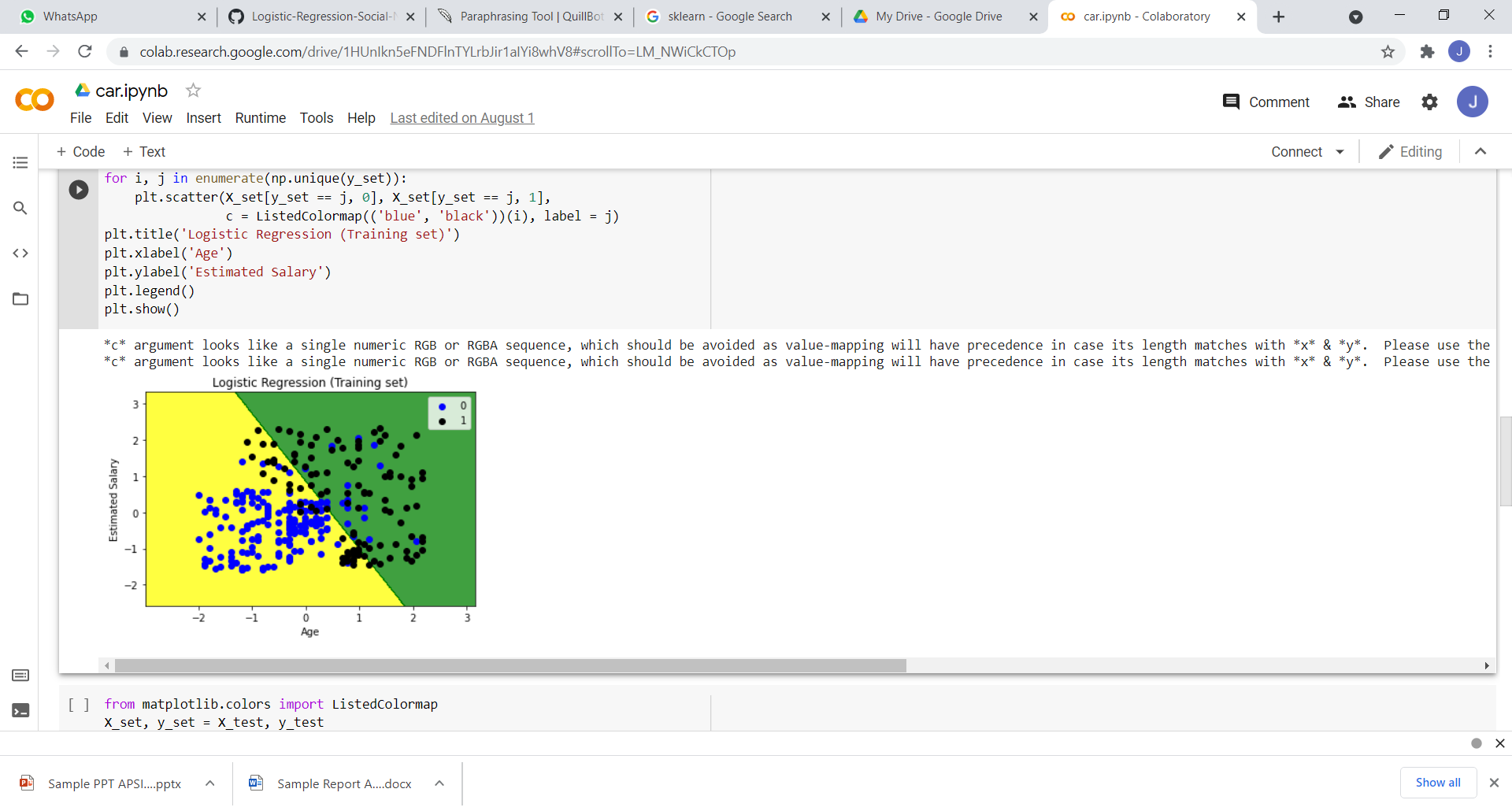
**About Dataset:** The dataset contains entries which contains the userId, gender, age, estimated salary and the purchased history. The matrix of features taken into account are age and estimated salary which are going to predict if the user is going to buy new car or not (1=Yes, 0=No).

**Solution:**  First the pre-processing of data is done and then the prediction was done using logistic regression followed with visualization. The confusion matrix below shows that our model predicts 89 correct and 11 wrong decisions which shows 89% accuracy.

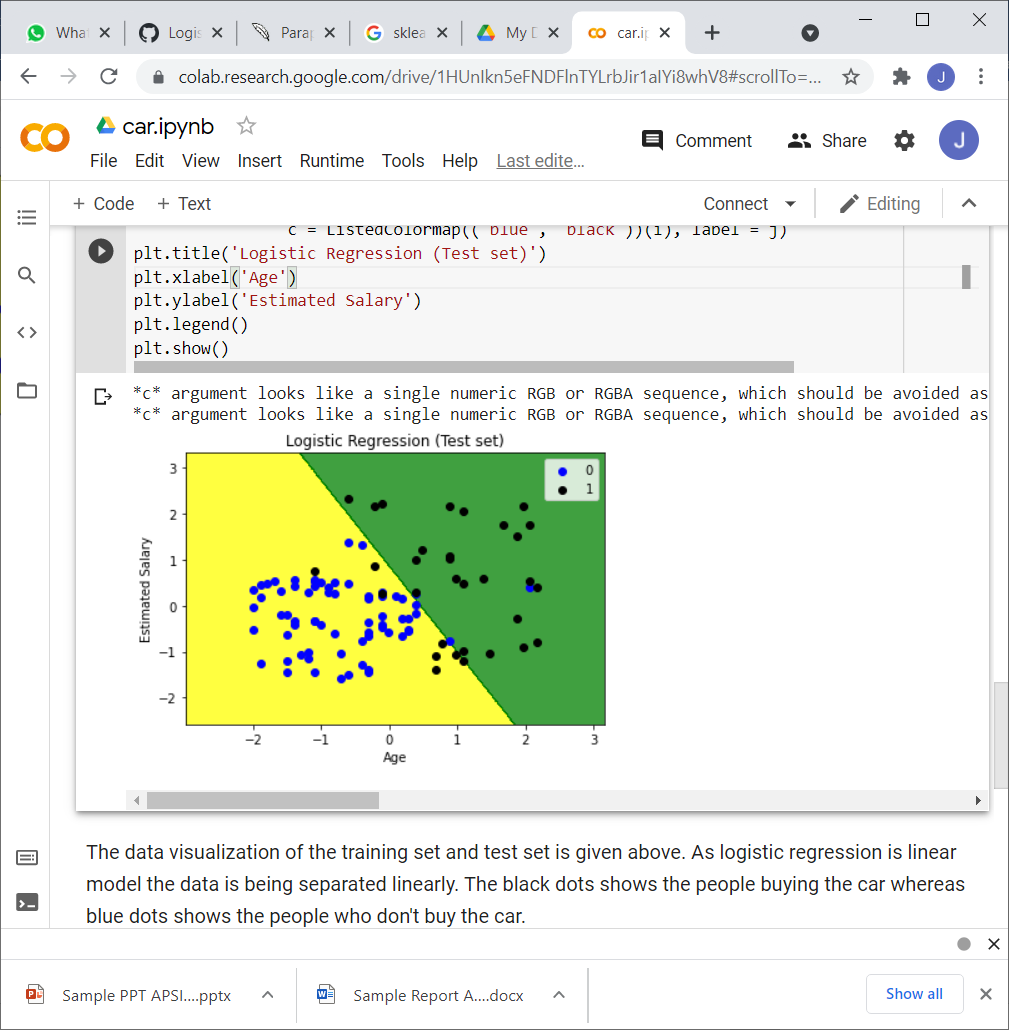


3

# Output



**Fig1 Training Set**



**Fig 2 Test Set**

**The data visualization of the training set and test set is given above. As logistic regression is**

**linear model the data is being separated linearly. The black dots show the people buying the car whereas blue dots shows the people who don't buy the car**.

4

# Applications

* Sales of Particular Products.
* Identify the buyer’s ratio.
* Marketing effects on Products.

5

**Advantages of using Logistic Regression**

* Logistic regression is easier to implement, interpret, and very efficient to train.
* It makes no assumptions about distributions of classes in feature space.
* Good accuracy for many simple data sets and it performs well when the dataset is linearly separable.
* It is very fast at classifying unknown records.
* It can interpret model coefficients as indicators of feature importance.

6

**Disadvantages of using Logistic Regression**

### If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting.

### The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables.

### Non-linear problems can’t be solved with logistic regression because it has a linear decision surface. Linearly separable data is rarely found in real-world scenarios.

7

# Conclusion

Thus, we have developed a project on Car Buying Prediction based on the dataset by using machine learning under this internship provided by Apsit Skills.

**References:**

[1] https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression

[2]https://colab.research.google.com/notebooks

8