DATA SCIENCE CASE STUDY:

AUTO FOLLOW SYSTEM OF

TESLA AUTOPILOT

# Project Description

## 1.1 What is Auto Follow System of Tesla Autopilot

As a new manufacturer of auto industry, Tesla has not only developed breakthrough technology for electric vehicle batteries, but also achieved great success in the field of autopilot (Eady, 2019). Autopilot is a cutting-edge technology of auto industry that with this technology, it is possible for human to liberate hands and let the car drive itself, for example in its auto following mode, it will automatically follow the front car and predict whether the front car is going to change the lane.

## 1.2 Introduction

To implement the autopilot system, a bunch of data science works needs to be done such as training data to make the system learn the real-time traffic and take appropriate measures and so on. This report focuses on the autopilot system of Tesla by analysing the data models and data technologies used in autopilot technology, exploring how these data be generated then finally make decision whether the front car is going change the lane in the following mode.

## 1.3 Data Science job roles

By collecting information from the official website of Tesla (<https://www.tesla.com/en_AU/careers/search#/?keyword=data>), and analysing the different aspects of the whole system, there are several significant data science job roles in this project.

### 1.3.1 Data Analyst

A Data Analyst take the responsibility to identify and improve infrastructure and processes of getting data. Ensure environments are stable and available. Except from these, a data analyst also needs to translate many complex technical issues into business terms and make optimized design to help machine learning scientists do better job.

### 1.3.2 Data Visualization Engineer

As a Data Visualization Engineer, the role needs to work within the Computer Vision Group and cooperate with machine learning scientists. This role will productize, code and finally test the Vision stack, improve its efficient and stability on generating data on a bunch of platforms. The data would later be calculated together with the image data translated from shooting of cameras.

### 1.3.3 Machine Learning Scientist

A Machine Learning Scientist takes the responsibility of research, design, optimize and implement the data models and algorithms. This role will use a computing cluster to train deep learning and machine learning models on detecting road environment and generalizing data segmentation to do analysing tasks. Besides, the role also needs to develop algorithms for multi-task learning, tracking, multi-sensor fusion and many other areas related to the autopilot technology.

### 1.3.4 Data Annotation Specialist

A Data Annotation Specialist needs to mark the images that critical to the deep neural networks by using the Autopilot labelling interface to label images for cars, lanes, street signs and cooperate with the Machine Learning Scientist to improve the design of an efficient labelling interface.

# Business Model

## 2.1 Business Application area

The Autopilot System is used in the condition of driving on the road, hence the conditions and decisions on the way should be taken into consideration. Figure 1 is the influence diagram of the Tesla Autopilot showing how the known challenges like signs, crossroads, day or night, type of road, available power etc. and unknown challenges like pedestrians, vehicles around etc. that affect the actions of the system.

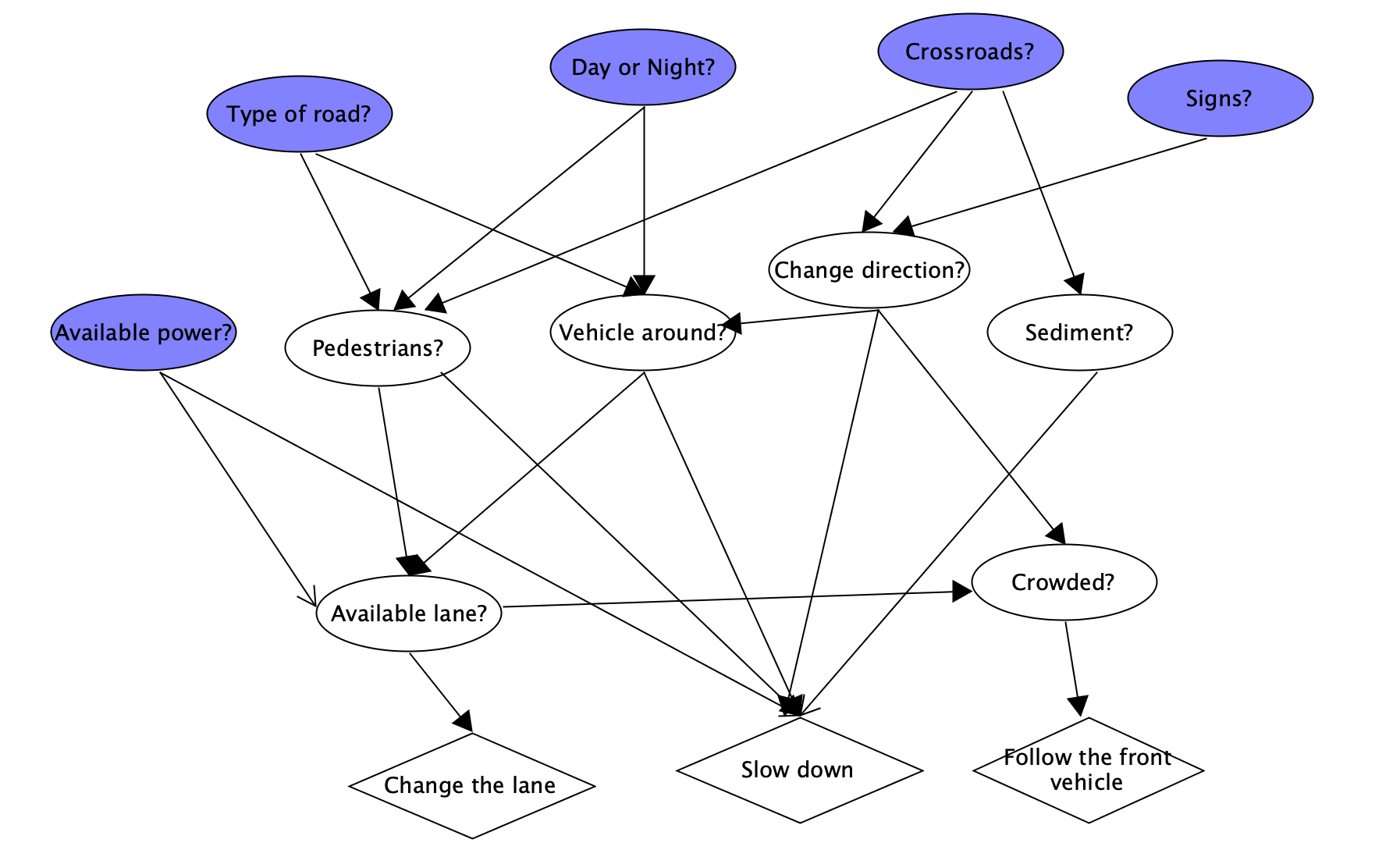


Figure Influence diagram for Autopilot System

## 2.2 Challenges

Since the condition of the road is quite complex and difficult to predict, how to get accurate real-time data is a great challenge. In order to handle such problem, several cameras and sensors are designed to be equipped around the surface of the vehicle. Through the cameras the whole environment near vehicle will be shoot and transferred into digital signals, together with the computer vision captured by tiny sensors, the real-time circumstances will be resolved as datasets and then calculated by the computing units inside the vehicle. After that, the predictions like whether the vehicles nearby would change lanes will be generated and the decision like slowing down and following the front vehicle would be made. Apart from that, how to design a training dataset and how to research out a machine learning algorithm with high efficiency and high accuracy is the most significant challenge for the system.

## 2.3 Values

The Autopilot System of Tesla is very meaningful and may cause subversive change to the auto industry. “The auto industry is not resistant to innovation and change but does tend to adapt slowly.” Kate Patrick (2018) said. This system will liberate drivers’ hands and improve driving pleasure. Further, human may make mistake while changing the lane during driving, but a well-trained car does not, so if this technology turns to maturity, the rate of traffic accidents will decline significantly.

# Characterising the Data and Data Processing

This section analyses the characteristics of the data that would be processed and how they will be processed using NIST analysis. These processed data will finally help Tesla Autopilot to implement the function of auto deciding whether the front car is going to change the lane.

## 3.1 Data characteristics

*Table 1* below summarizes the characteristics of data and data processing.

Table : Data characteristics

|  |  |
| --- | --- |
| Data Characteristics | Description |
| Data sources | There are several tiny sensors set around the car, these sensors detect the environments nearby and then the data will be collected and generated by the CPU of the car. |
| Data volume | In order to have better accuracy of predicting whether it is suitable to change the lane as human does, plenty of training and testing data are needed, so 100 gigabytes to 10 terabytes would be a reasonable numerical interval. |
| Data velocity | Since the circumstances change continuously, e.g. the number of cars around, crossroads, the condition of front car etc., so the data would be detected by the sensors each second. |
| Data variety | * Position data of front car * Number of cars around * Data of current time * Data of road condition * Data of weathers |
| Data veracity | The front car may make wrong decision to change the lane and cause an accident. |
| Software | Pandas, Spark |
| Analytics | Auto Follow System |
| Processing | CPU, tiny sensors |
| Lifecycle | Continued detecting and predicting |

## Required Technologies for data processing

### 3.2.1 Python

Since the data generated by the CPU of the car will finally be csv files, Python is a good choice as programming language to process the data due to its clearly gramma, strong performance on data and abundant machine learning API libraries.

### 3.2.2 Pandas

Pandas is a technical software library designed for Python to process tables and use the tables to plot ideal graphs, it will make the visualisation more elegant and professional.

### 3.2.3 Spark

In order to do the most important section, machine learning, spark is chosen as a suitable technology for its outstanding performance on machine learning.

# Resources

## 4.1 Gather Data

After searching through several open-dataset websites, it can be conjectured that Tesla does not release its training and testing dataset of Auto Follow System to the public, so I make the decision to generate a dummy dataset as an alternate to do the data analysis and test the result of machine learning. The dataset which contains more than 100 records can be attached through https://drive.google.com/open?id=1cR7zYEU8iMUgJvPEmYEEyfl8lmB\_YYJh.

## 4.2 Data Exploration

In the dummy dataset, we have 7 columns: Day\_Or\_Night, Cars\_Around, Windy, Rainy, Position1, Position2, Is\_Front\_Car\_Change\_Lane. The dataset is processed via Python, Pandas and Spark. As shown in Figure 2.

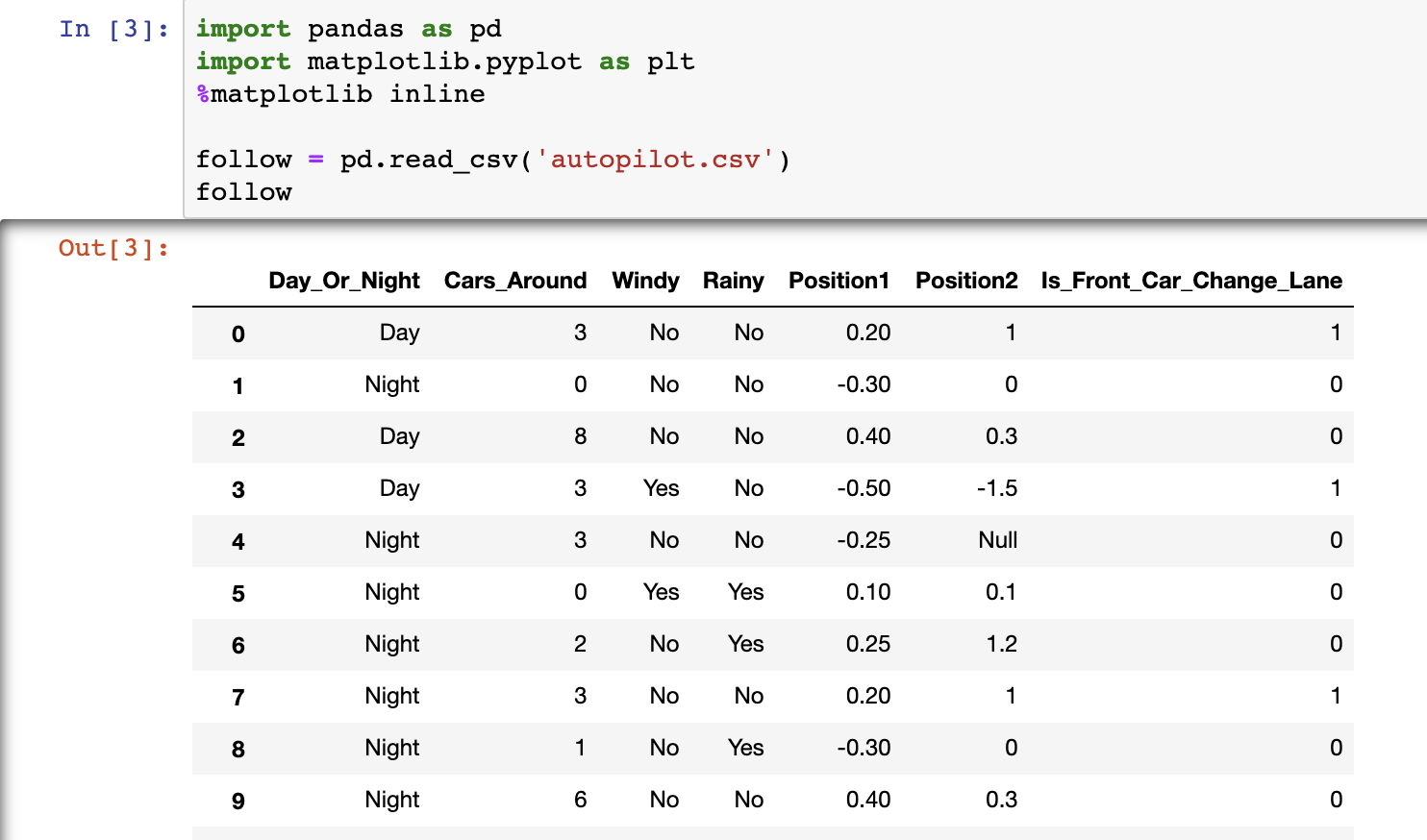


Figure Columns

The ‘Day\_Or\_Night’ column records the time that the data is detected (Day or Night), this would be a factor to affect the decision of drivers to change a lane since people have different behaviours in day and night.

The ‘Cars\_Around’ column records the cars driving on the road around Tesla car. Since changing a lane should take the driving status of other cars on the road into consideration, the amounts of cars around is an important indicator.

The ‘Windy’ column records the wind condition during driving because strong winds may cause the vehicle to slip, which affects the driver’s tendency to change lanes.

The ‘Rainy’ column performs similar functions as ‘Windy’ because rainy days can reduce visibility and can cause tires to slip, which are dangerous signals for drivers on the road and may affect a driver’s decision.

The ‘Position1’ and ‘Position2’ record two relative positions of the Tesla vehicle and front car in a second, negative digits mean the front car is on the left front of the Tesla vehicle and positive digits means the front car is on the right front. Apart from these, the digits mean the relative distance (e.g. ‘-0.5’ means the front car is 0.5 meters in front of the left). If the two positions are far apart, that represents that the front car has moved a lot of distance horizontally, which would be an obvious judgment basis of whether the driver tends to change the lane.

The final column is ‘Is\_Front\_Car\_Change\_Lane’, which is a column that records the results of whether the driver finally change a lane (1 means change the lane and 0 means not change the lane). This is a necessary column to test the accuracy of the predictions made by machine learning algorithms.

### 4.2.1 Descriptive analysis

Through the columns above, many analyses can be implemented. For example, a comparison of whether the front cars change the lanes can be done in conditions of when the time is day and night. As shown below in Figure 3.



Figure Comparison of changes in day and night

From the plot the different changing amounts in day and night are clearly shown, the amount in day is 30 and that in night is 20, thus we can speculate that when the time is daytime, the front car has a greater possibility of lane change.

Similarly, a comparison of lane change in different weathers is shown in Figure 4.



Figure Weather

In order to fix such errors, a pre-processing to the data is needed.

## 4.3 Data Pre-processing

To do machine learning, the dataset should be firstly loaded into spark context, and the abnormal data in the dataset should firstly be handled, such as the nulls in dataset (As shown in Figure 5).

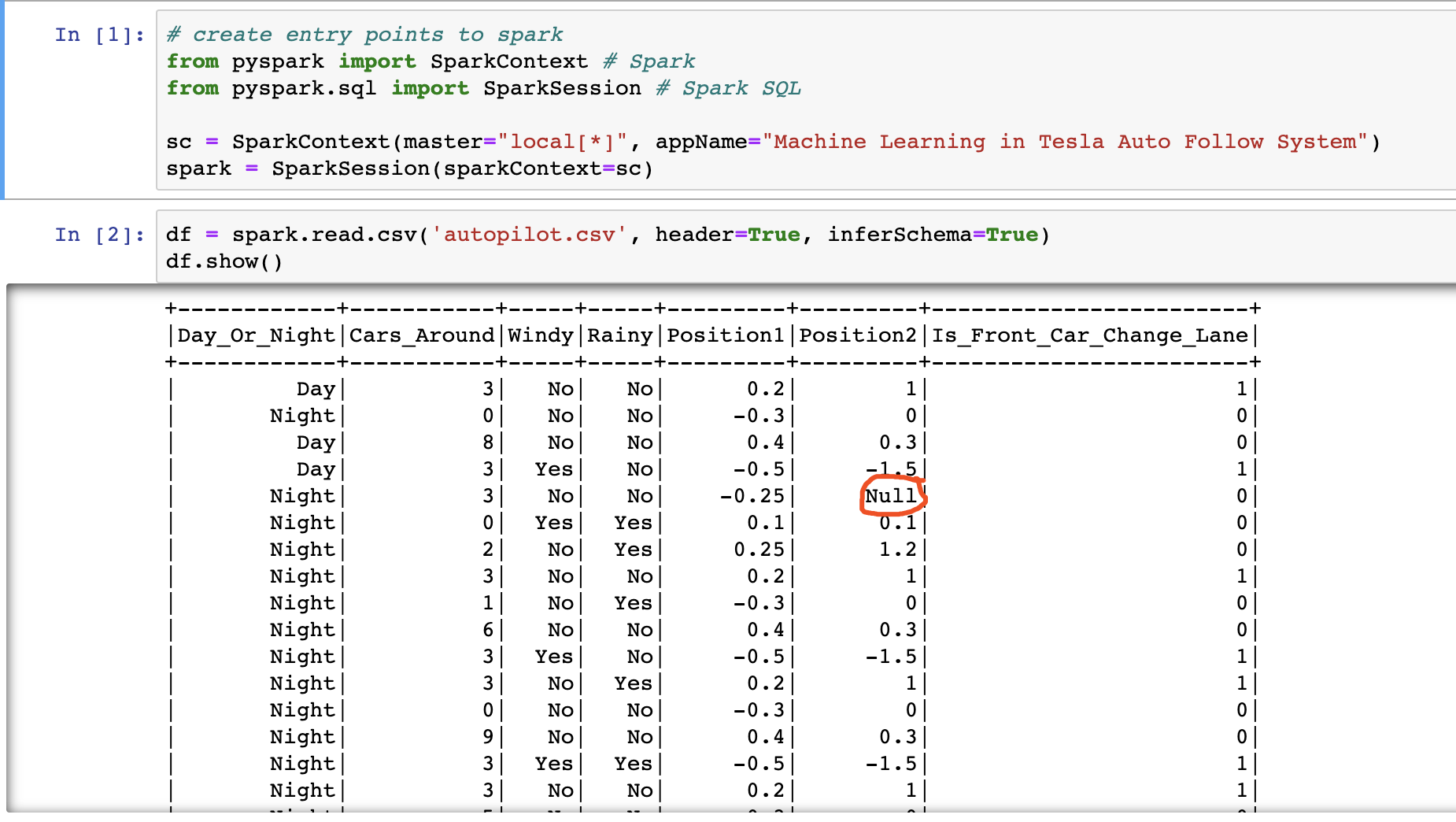


Figure Null data in dataset

Since the column that contains nulls is a numerical column, the nulls will be replaced by the average digit of ‘Position2’. As shown in Figure 6, the Null value has been successfully replaced.

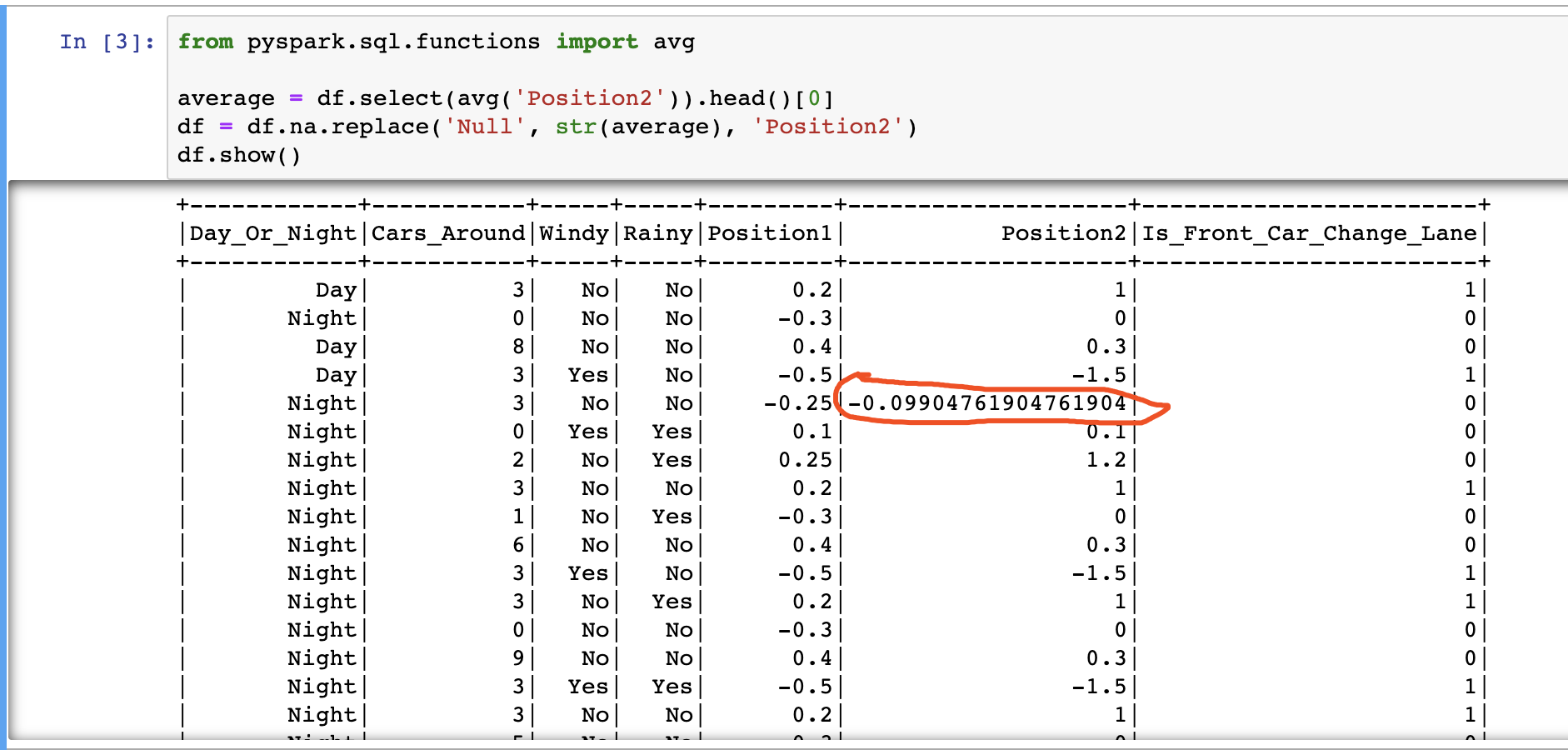


Figure Replace Null Values

There is a problem remaining in 4.2.1 that it is not a numerical column. Because the data type of columns ‘Windy’ and ‘Rainy’ are non-numerically, would be. In this condition, some transfer to the data type is needed. One-Hot encoding is a good way to do this transform.

One-hot encoding, also known as one-bit efficient encoding, uses an N-bit status register to encode N states, each state having its own separate register bits, and at any one time, only one of them is valid.

Follow this encoding method we can consider the data ‘No’ as 0, and ‘Yes’ as 1, so the visualization could be done, and this will help machine learning algorithms run more efficiently. To better implement One-hot encoding, a Python class called StringIndexer could be used to transfer non-numerical columns to numerical columns. Also, since the original data type of columns is string, the numerical columns should be transferred to double type, as shown below.

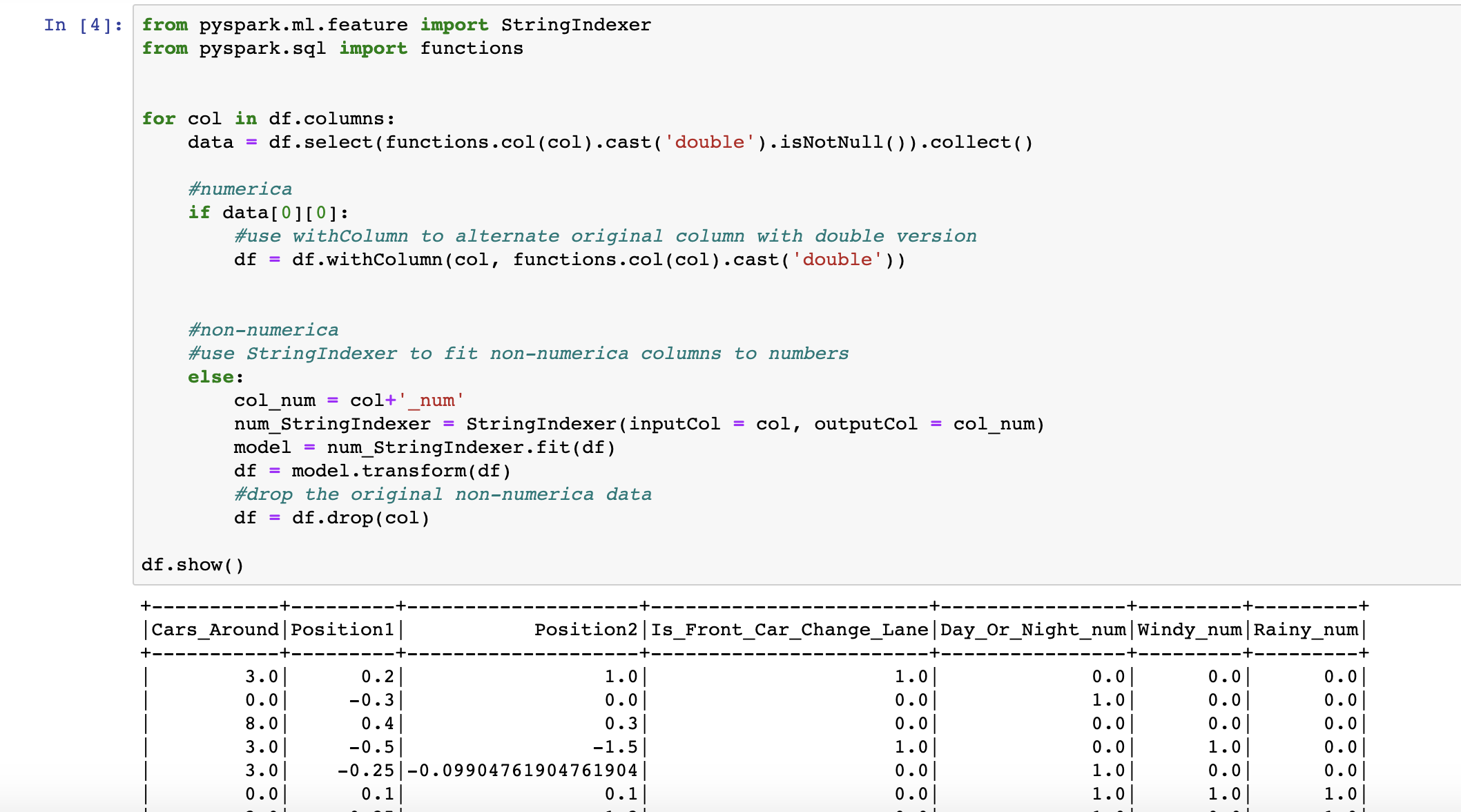


Figure Transfer data types

After the transform of columns, the finally pre-processing step is to create the feature vector and divide the dataset into training data and test data. VectorAssembler is used here to assemble the columns into a feature vector. (‘Is\_Front\_Car\_Change\_Lane’ is excluded here in order to test the accuracy of predictions after training.) The code of pre-procession and the output is shown in Figure 8.

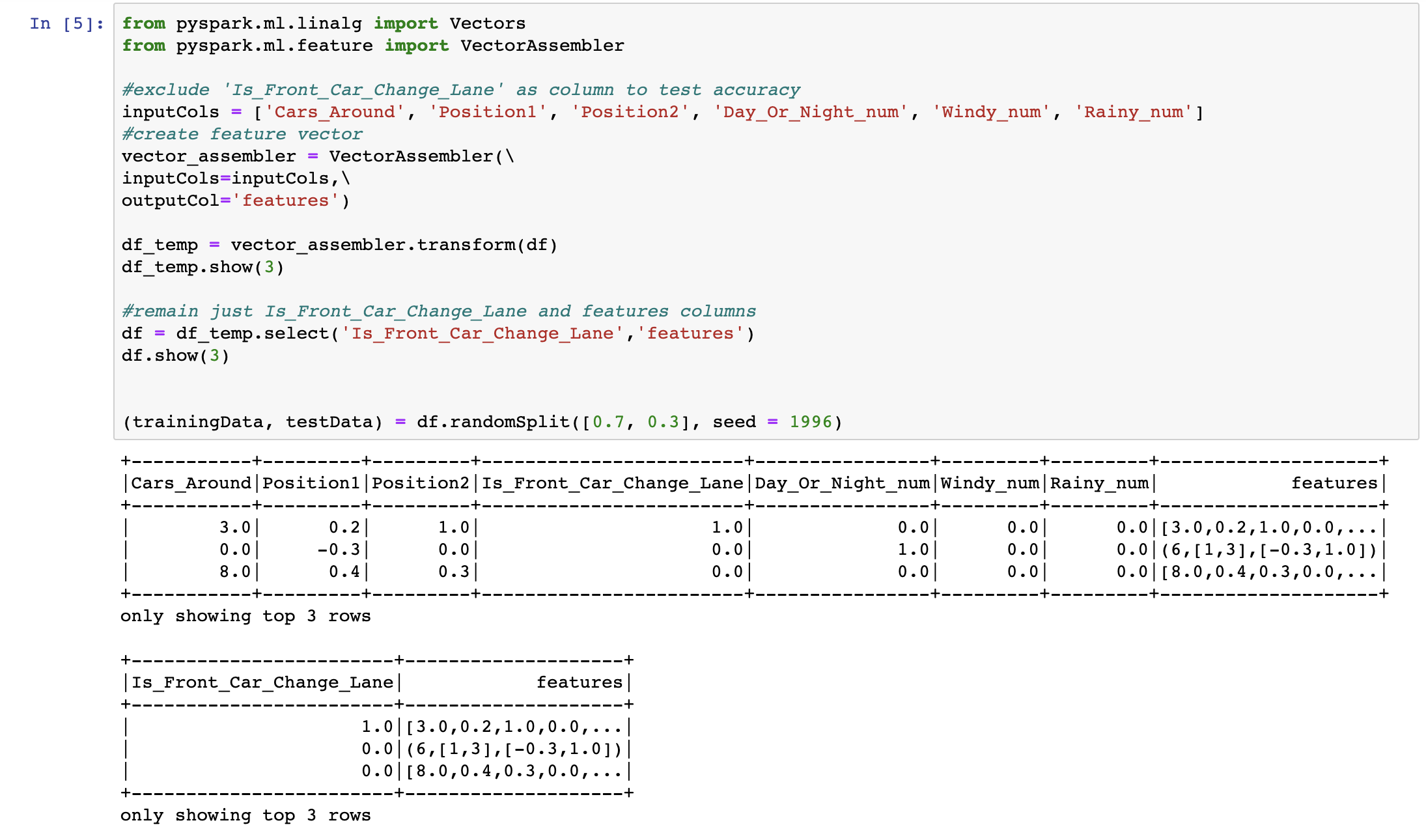


Figure Create the feature vector and divide the dataset

# Data Analysis

## 5.1 Machine Learning Algorithms

In order to do the machine learning, a suitable algorithm is very important, in this part, three machine learning algorithms would be taken into consideration.

### 5.1.1 Decision Tree

The decision tree is a relatively simple one in the machine learning supervised learning classification algorithm. The decision tree is a predictive model; it represents a mapping relationship between object attributes and object values. Each node in the tree represents an object, and each forked path represents a possible attribute value, and each leaf node corresponds to the object represented by the path from the root node to the leaf node.

### 5.1.2 Logistic Regression

The model of logistic regression is a nonlinear model, also known as a logistic regression function. But it is essentially supported by linear regression.

### 5.1.3 Gradient Boosted Tree

GBT is a widely used algorithm that can be used for classification and regression. Has a good effect on a lot of data.

5.2 Comparison of Machine Learning Algorithms

In this part the three machine learning algorithms are implemented to train the dataset, the accuracies of them will be compared and the algorithm with better performance on accuracy will be finally chosen.

The implementation and output of Decision Tree is shown below.

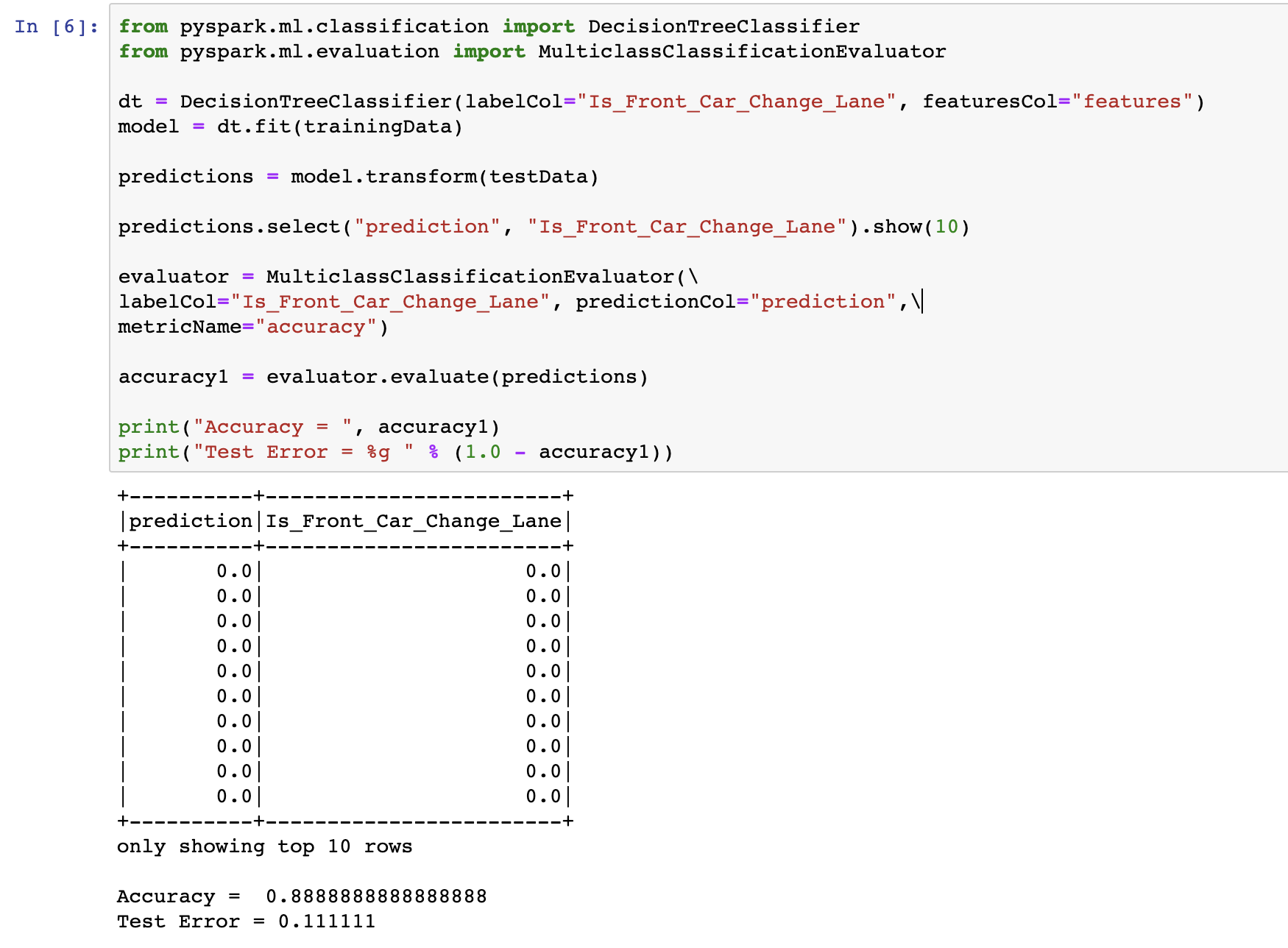


Figure Decision Tree

The implementation and output of Logistic Regression is shown below.



Figure Logistic Regression

The implementation and output of Gradient Boosted Tree is shown below.

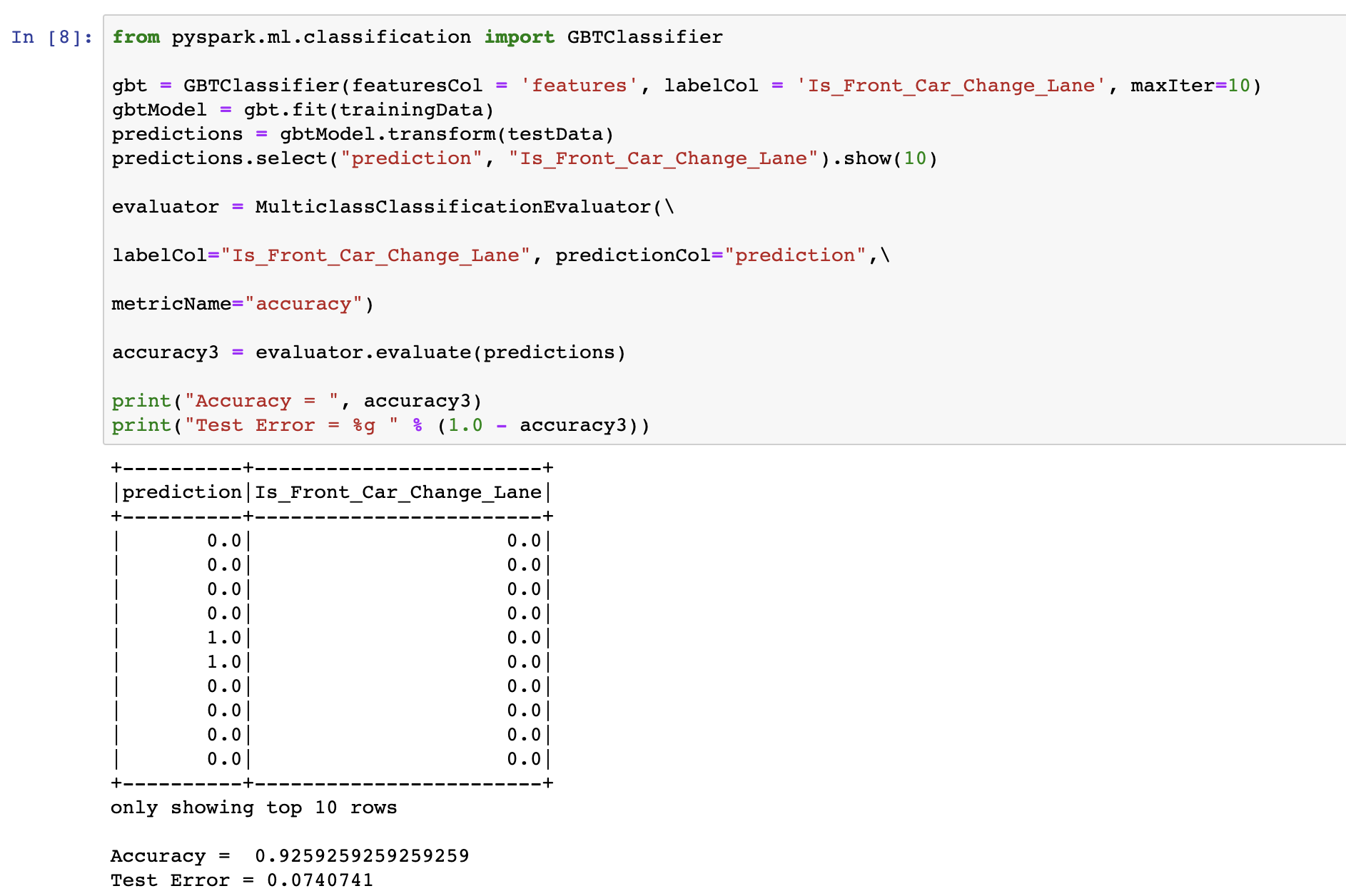


Figure Gradient Boosted Tree

After calculating the accuracies of the three algorithms, a bar chart is plotted to compare them.

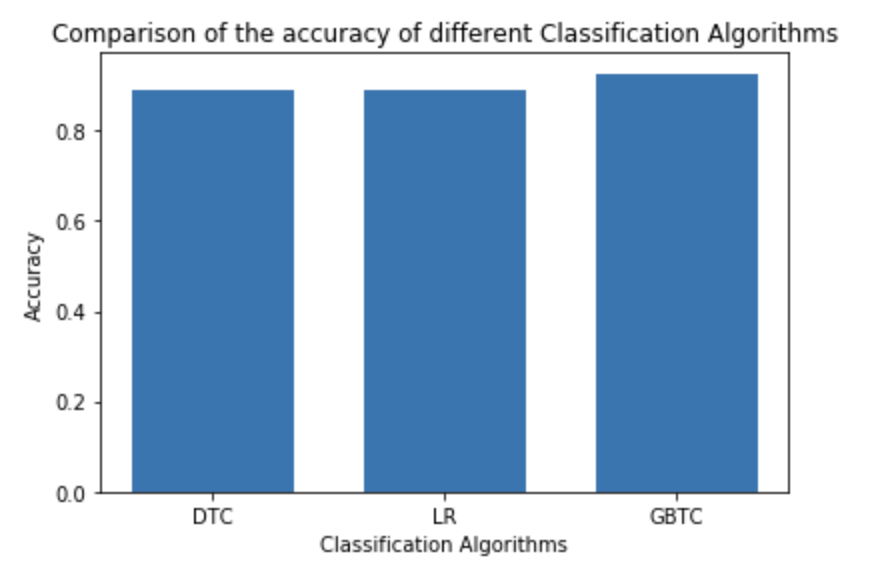


Figure Comparison of the accuracy of different Classification Algorithms

From the graph it is clearly that Gradient Boosted Tree has better accuracy to predict whether the front car is going to change the lane, so GBT is the best choice to perform as a machine learning algorithm in Auto Follow System of Tesla.

# Conclusion

After complex and accurate data analysis, Tesla's Auto Follow System implementation process and the most appropriate algorithm were successfully simulated in this case.

# Reference

Patrick, K. (2019). *Case study: How Tesla changed the auto industry*. [online] Supply Chain Dive. Available at: https://www.supplychaindive.com/news/case-study-how-tesla-changed-the-auto-industry/517251/ [Accessed 17 Oct. 2019].

Eady, T. (2019). *Tesla’s Deep Learning at Scale: Using Billions of Miles to Train Neural Networks*. [online] Medium. Available at: https://towardsdatascience.com/teslas-deep-learning-at-scale-7eed85b235d3 [Accessed 7 May 2019].