Table of Content

Introduction

Problem Statement

With the advent of big data and machine learning, it has been more efficient than ever to gain insight into numerous things that had little correlation before. This requires data that is gathered from a controlled group and the likelihood of them purchasing a product can be affected by a number of factors. But fortunately 90% of the world's data was generated in the past two years only. Machine learning aims to gather all these data and factors to extract value. In this report, we compare and contrast two different algorithms that predicts the possibility of someone purchasing an iPhone based on their age and salary.

In our huge consumer based product market, predicting the likelihood of purchase for a product is essential to ensure the product's success. Traditionally this is done through surveys and product testing in a closed environment, which is not cost effective and takes away time that the company could use to actually sell. Statistical and anecdotal methods are used to conduct social research to optimise the sale of any particular product. The actual term for this kind of research and information gathering is called market analysis. It is also done to investigate the demographics of the consumers who are more inclined to purchase the product, if a company needs to attract more consumers by adding features or niche down to target specific people. A source of data, for example for a company like Apple, is the feedback from previous product its success. This reduces overall costs of launching a product and an estimate can be done to see if the investment would be fruitful before committing into developing a product. This is why we see sometimes products are not sold or developed anymore such as the iMac but are then re-produced because there was a demand.

Objective

Our objective in this group project is to predict if the customer will purchase an iPhone or not given their gender, age and salary.

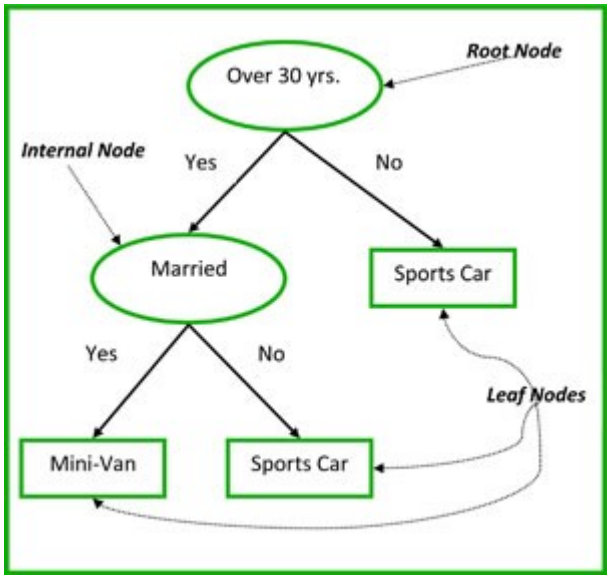
Learning Algorithm

Decision Tree Algorithm

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

A decision tree is a flowchart-like structure in which each internal node represents a “test” on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

Tree based learning algorithms are considered to be one of the best and mostly used supervised learning methods. Tree based methods empower predictive models with high accuracy, stability and ease of interpretation. Unlike linear models, they map non-linear relationships quite well. They are adaptable at solving any kind of problem at hand (classification or regression).



Example: Decision Tree

Common terms used with Decision trees:

Root Node: It represents entire population or sample and this further gets divided into two or more homogeneous sets.

Splitting: It is a process of dividing a node into two or more sub-nodes.

Decision Node: When a sub-node splits into further sub-nodes, then it is called decision node.

Leaf/ Terminal Node: Nodes do not split is called Leaf or Terminal node.

Pruning: When we remove sub-nodes of a decision node, this process is called pruning. You can say opposite process of splitting.

Branch / Sub-Tree: A subsection of entire tree is called branch or sub-tree.

Parent and Child Node: A node, which is divided into sub-nodes is called parent node of sub-nodes whereas sub-nodes are the child of parent node.

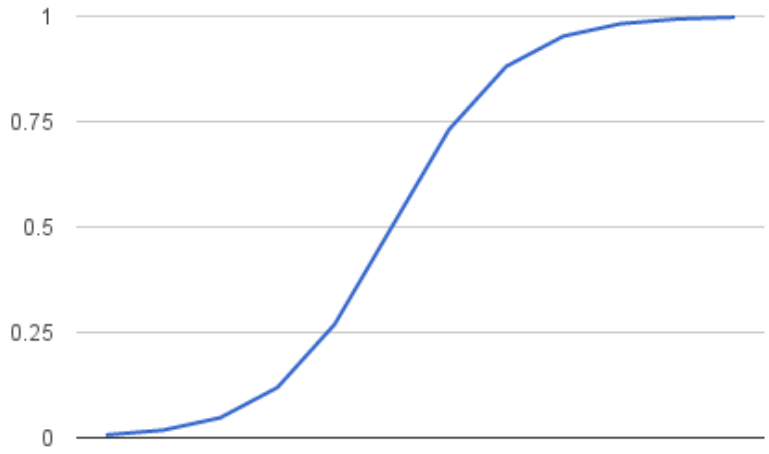
Logistic Regression Algorithm

Logistic regression is named for the function used at the core of the method, the logistic function.

The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

$$1 / (1 + e^{-value})$$

Where e is the base of the natural logarithms (Euler's number or the EXP() function in your spreadsheet) and value is the actual numerical value that you want to transform. Below is a plot of the numbers between -5 and 5 transformed into the range 0 and 1 using the logistic function.



Example: Logistic Function (Sigma Function Graph)

Representation Used for Logistic Regression

Logistic regression uses an equation as the representation, very much like linear regression.

Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modeled is a binary values (0 or 1) rather than a numeric value.

Below is an example logistic regression equation:

$$y = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)})$$

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.

The actual representation of the model that you would store in memory or in a file are the coefficients in the equation (the beta value or b's).

Methodology

Logistic Regression

```python
# Step 1 - Load Data
import pandas as pd
dataset = pd.read_csv("iphone_purchase_records.csv")
X = dataset.iloc[:,:-1].values
y = dataset.iloc[:, 3].values

# Step 2 - Convert Gender to number
from sklearn.preprocessing import LabelEncoder
labelEncoder_gender =  LabelEncoder()
X[:,0] = labelEncoder_gender.fit_transform(X[:,0])

# Step 3 - Split Data into training and testing
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25,random_state=0)

# Step 4 - Feature Scaling
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)

# Step 5 - Logistic Regression Classifier
from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression(random_state=0, solver="liblinear")
classifier.fit(X_train, y_train)

# Step 6 - Predict
y_pred = classifier.predict(X_test)

# Step 7 - Confusion Matrix
from sklearn import metrics
cm = metrics.confusion_matrix(y_test, y_pred)
print(cm)
accuracy = metrics.accuracy_score(y_test, y_pred)
print("Accuracy score:",accuracy)
precision = metrics.precision_score(y_test, y_pred)
print("Precision score:",precision)
recall = metrics.recall_score(y_test, y_pred)
print("Recall score:",recall)
```

Decision Tree

```python
# Step 1 - Load Data
import pandas as pd
dataset = pd.read_csv("iphone_purchase_records.csv")
X = dataset.iloc[:,:-1].values
y = dataset.iloc[:, 3].values

# Step 2 - Convert Gender to number
from sklearn.preprocessing import LabelEncoder
labelEncoder_gender =  LabelEncoder()
X[:,0] = labelEncoder_gender.fit_transform(X[:,0])

# Step 3 - Split Data
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=0)

# Step 4 - Fit the classifier
from sklearn.tree import DecisionTreeClassifier
classifier = DecisionTreeClassifier(criterion = "entropy", random_state=0)
classifier.fit(X_train, y_train)

# Step 5 - Predict
y_pred = classifier.predict(X_test)

# Step 6 - Evaluate the model performance
from sklearn import metrics
cm = metrics.confusion_matrix(y_test, y_pred)
print(cm)
accuracy = metrics.accuracy_score(y_test, y_pred)
print("Accuracy score:",accuracy)
precision = metrics.precision_score(y_test, y_pred)
print("Precision score:",precision)
recall = metrics.recall_score(y_test, y_pred)
print("Recall score:",recall)
```

Results

<table>
<tr><td colspan="2">Logistic Regression</td><td>Vs</td><td colspan="2">Tree Decision</td></tr>
<tr><td colspan="2"><table>
<tr><td colspan="2">Confusion matric</td><td>Predicted</td><td></td></tr>
<tr><td colspan="2"></td><td>Negative</td><td>Positive</td></tr>
<tr><td rowspan="2">Actual</td><td>Negative</td><td>65 (TN)</td><td>3 (FP)</td></tr>
<tr><td>Positive</td><td>6(FN)</td><td>26 (TP)</td></tr>
</table></td><td>Confusion Matrix</td><td colspan="2"><table>
<tr><td colspan="2">Confusion matric</td><td>Predicted</td><td></td></tr>
<tr><td colspan="2"></td><td>Negative</td><td>Positive</td></tr>
<tr><td rowspan="2">Actual</td><td>Negative</td><td>63 (TN)</td><td>5 (FP)</td></tr>
<tr><td>Positive</td><td>3FN)</td><td>29 (TP)</td></tr>
</table></td></tr>
<tr><td colspan="2">0.91</td><td>Accuracy score</td><td colspan="2">0.92</td></tr>
<tr><td colspan="2">0.896551724137931</td><td>Precision score</td><td colspan="2">0.8529411764705882</td></tr>
<tr><td colspan="2">0.8125</td><td>Recall score</td><td colspan="2">0.90625</td></tr>
</table>

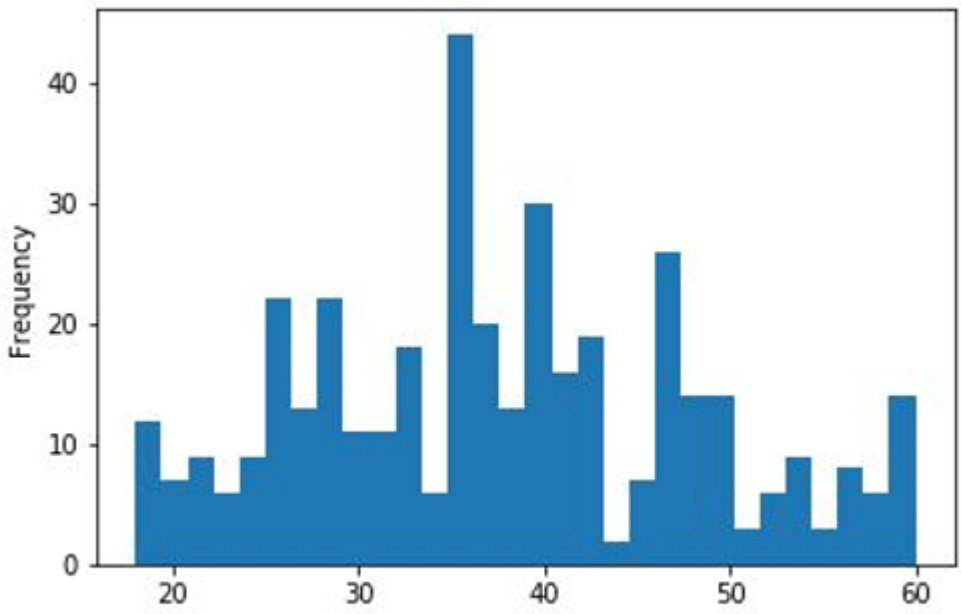
Accuracy Score = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall Score = TP / (TP + FN)

Data Visualization

Histogram Of Respondent's Age

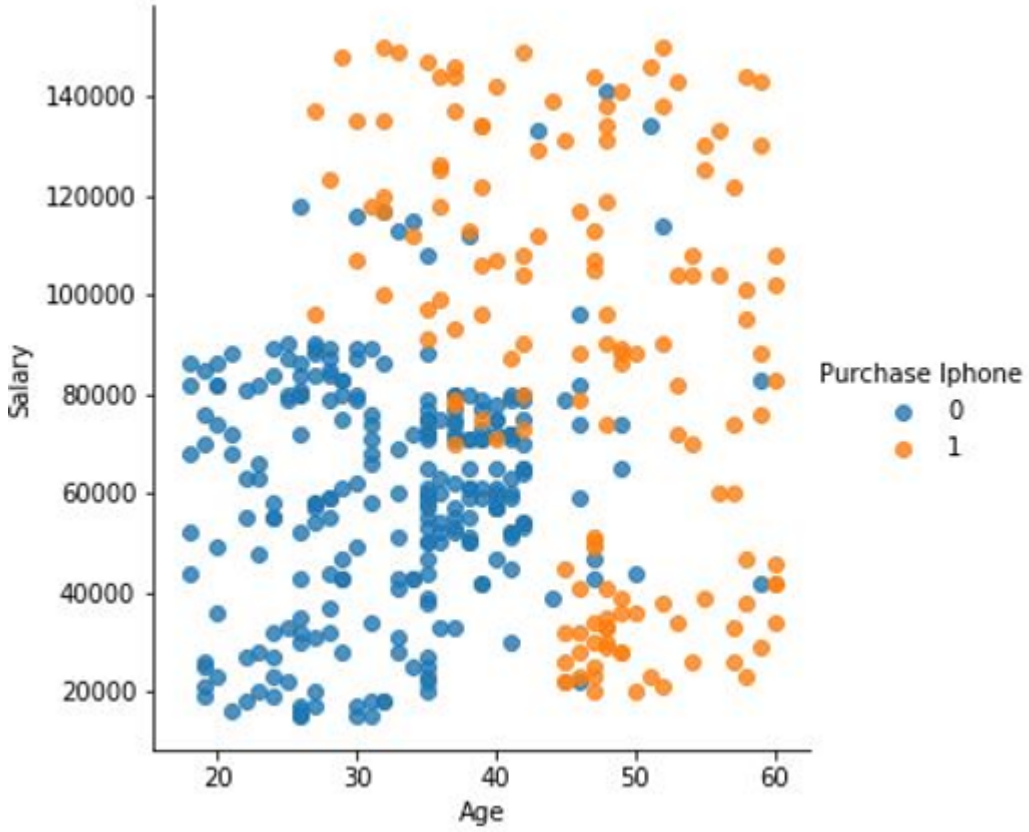


Histogram of Respondent's Age

A histogram is a graphical display of data using bars of different heights. In a histogram, each bar groups numbers into ranges. Taller bars show that more data falls in that range. A histogram displays the shape and spread of continuous sample data.

In the histogram of respondent's age it is found that respondent from age 20 to 40 has bought iphones where the respondent whose age is around 35 has bought around 40 iphones.

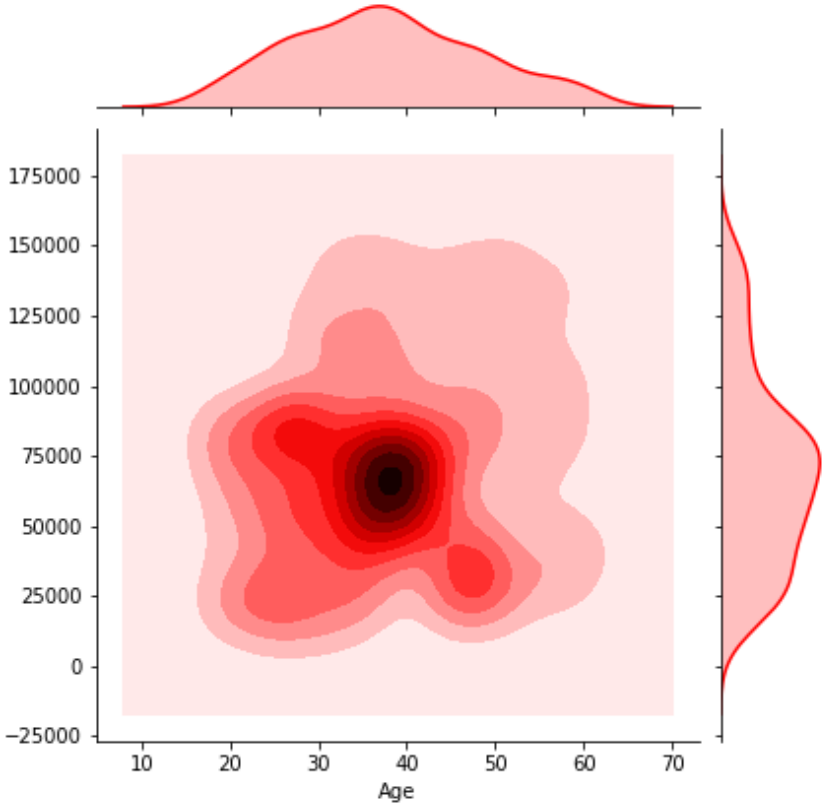
Scatter Plot of Salary and Age of Respondent that purchase Iphone



Scatter Plot of Salary and Age of Respondent that purchase Iphone

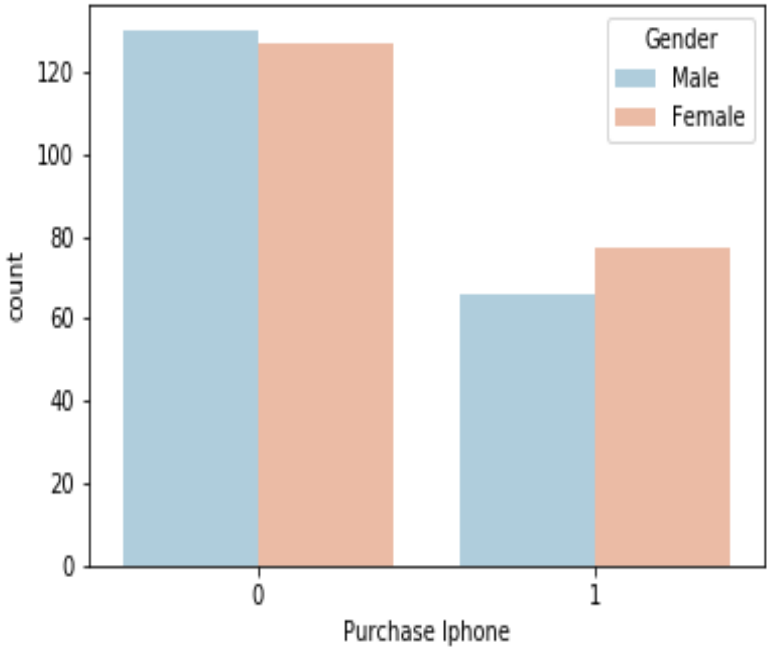
A scatter plot (also called a scatter plot, scatter graph, scatter chart, scattergram, or scatter diagram) is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data.

Joint Plot of Salary and Age of Respondent



Joint Plot of Salary and Age of Respondent

A marginal plot allows to study the relationship between 2 numeric variables. The central chart display their correlation. It is usually a scatter plot, a hexbin plot, a 2D histogram or a 2D density plot. ... The seaborn library provides a joint plot function that is really handy to make this type of graphic.



Amount of respondent that purchase and not purchase iphone according to gender

Conclusion

The technology is moving towards high performance and personalized content which catcher accordingly to its user. Machine Learning algorithm will play a major role in this change, as computer model processes data (big data) way faster compared to humans.

Relatively each algorithm is suited best to solve a certain problem based on the time complexity, data size and results.

References

1. Ayush Pant, 2018 - Introduction to Logistic Regression Retrieved from : https://towardsdatascience.com/introduction-to-logistic-regression66248243c148

2. Rajesh S. Brid, 2018 - Decision Tree - A Simple Way to Visualize a Decision Retrieved from : https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb

3. Jason Brownlee, 2016 - Logistic Regression for Machine Learning Retrieved from : https://machinelearningmastery.com/logistic-regression-for-machine-learning/