Tampere University of Applied Sciences

Final Project

Machine Learning

Md Touhidul Islam

BACHELOR’S
Month 2021

Degree Programme
Machine learning

ABSTRACT

Tampereen ammattikorkeakoulu
Tampere University of Applied Sciences
Name of the Degree Programme
Software Engineering

Bachelor's thesis xxs pages, appendices x pages
Month 2020

Key words: please use lower case initial letters

CONTENTS

1 INTRODUCTION

In this project going to work with mall customer’s data. It has 200 rows and 5 columns. The columns in this dataset are CustomerID, Gender, Age, annual income and Spending score. Most of the calculation will do using age and annual income of the customer. First took a variable that name is customer variable. In this variable wrote the code for reading the csv file. Checked the first five and last five data using head and tail function. Then describe function showing the mean, std, max and etc of data. Shape showing the amount of row and columns. Also, isnull function showing the null value that not available in data sheet. Using heatmap got the correlation between columns. Strongest corelation between Annual income and customer id. Lowest correlation between spending score and age. Pair plot method showing all correlation with a graph. Relplot method showing relation between annual income and age. From the beginning and end high income person was Female. Overall, the highest and lowest annual income person is male. Goal to calculate mean, accuracy for the data from different method. Must see which accuracy is better.

2 IMPLEMANTATION AND RESULTS

2.1 Data explore and visualize

After reading the csv file using the customer variable started exploring the data. First, customer.head(), customer.tail() show first and last five row of data. Then describe () show count, mean, std, min, 25%, 50%, 75%, max of data and shape() show the whole shape of data (200,5). Columns show all columns. Isnull() showing all null value but in this dataset null value not available. Corelation variable showing relationship. From the relationship heatmap got the strongest relation between Annual income and customer id and lowest relation between age and spending score. Pairplot show the relationship graph for understanding. Also, showing implot and relplot for annual income and age.

2.1.1 Regression and data pre-process

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables). In this dataset has one string that is Genre(Gender). For making our calculation easy we replace this columns attributes for female 0 and male 1.

```
customer['Genre'] = customer['Genre'].replace({'Female': 0, 'Male': 1})
```

then define x and y:

```
x = customer.iloc[:, :-1].values
y = customer.iloc[:, 1].values
```

Then start train and test:

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 25, random_state = 42)
```

used LinearRegression:

```
regressor = LinearRegression()
regressor.fit(x_train, y_train)
```

Calculated:

```
print('Coefficients: \n', regressor.coef_)
# The mean squared error
print("Mean squared error: %.2f" % np.mean((regressor.predict(x_test) - y_test) ** 2))
# Explained variance score: 1 is perfect prediction
print('Variance score: %.2f' % regressor.score(x_test, y_test))

print ('Mean: %.2f' % y_test.mean())
```

2.1.2 Decision tree

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. Defined x and y like regression and fit it then got the decision tree.

Got here:
Length of data, mean and accuracy.

2.2 Classification

Classification is the process of identifying and and grouping objects or ideas into predetermined categories. Classification enables the separation and sorting of data according to set requirements for various business or personal objectives.
Here, Took feature_cols and then define x and y.

Code:

```
feature_cols = ['CustomerID', 'Spending Score (1-100)', 'Age', 'Annual-IncomeKeuro']
x = customer[feature_cols]
y = customer.Genre
logreg = LogisticRegression()

train and test:
x_train, x_test, y_train, y_test = train_test_split(x,y, test_size=0.25)
logreg.fit(x_train, y_train)
prediction:
y_pred_class = logreg.predict(x_test)
print(metrics.accuracy_score(y_test, y_pred_class))
y_test.mean()
```

2.2.1 Result

Regression:

From the whole dataset:

Coefficients: [2.74904127e-03 -2.78491707e-03 -7.77266274e-05 -5.16357548e-03]

Mean squared error: 0.28

Variance score: -0.12

Mean: 0.50

Decision tree:

Length of x_train: 150

Length of y_test: 50

Mean: 1

Accuracy: 0.72 or 72%

Classification:

Length of x_train: 150

Length of y_test: 50

Accuracy: 0.48 or 48%, Mean: 0.5

3 Discussion

For method train and test length same. Regression and classification maen same but decision tree is different. Accuracy for tree based method is 72% but classification accuracy 48%. Here, Tree based methods accuracy better. For this dataset looks like tree based method is better. Suppose we thake, Method 1, Method 2, Method 3 like regression, decision tree, classification.
Method 1 mean = .5
Method 2 mean = 1
Method 3 mean = .5
Method two is better than other method.

Method 1 accuracy = Not much better
Method 2 accuracy = .72
Method 3 accuracy = .48

Method 2 more accurate then other method. If someone use method 2 for this type of dataset, it’s possible to get well accuracy.
All method doesn’t fix for all datasets. That’s the reason testing the accuracy using different method. It reached near to the better method.

4 Conclusions

It very tough to make a conclusion for above method. Accuracy and mean are change always. This project gone well from the beginning. Invested huge time for understanding the code and method though still not clear all concept. Still learning from this course. All method is good. For finding the best one have to analysis more deep.

5. REFERENCES

Semtu. VEMO-valuankkurit. Manual. Read 11.12.2021. https://en.wikipedia.org/wiki/Regression_toward_the_mean

Reference: Read 12/12/2021 http://www.scielo.org.za/scielo.php?script=sci_arttext&pid=S0011-85162016000600009

Reference: Read 12/12/2021 https://www.r-bloggers.com/2020/08/accuracy-of-forecasting-methods-can-you-tell-the-difference/

Reference: Read 12/12/2021 https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/

Reference: Read 12/12/2021 https://www.quora.com/My-model-train-accuracy-is-75-and-test-accuracy-is-70-is-it-overfitting

Reference: Read 11/12/2021 https://en.wikipedia.org/wiki/Mean

Reference: See 10/11/2021 https://www.youtube.com/watch?v=7eh4d6sabA0&ab_channel=ProgrammingwithMosh

6. APPENDICES

6.1 Appendix 1. Clean Jupyter notebook

https://drive.google.com/file/d/1mKh9iMboPuyI0LkDPG5WUFjsVls-VAtd/view?usp=sharing