**Table of the content**

**List of the tables**

**List of the figures**

|  |
| --- |
| **Problem 1**: Linear Regression |
| **1 Introduction** |  |
| * 1. Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA, duplicate values). Perform Univariate and Bivariate Analysis. | **6-12** |
| * 1. Impute null values if present, also check for the values which are equal to zero. Do they have any meaning, or do we need to change them or drop them? Check for the possibility of combining the sub levels of a ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning | **13-14** |
| * 1. Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from stats model. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning. | **14-20** |
| * 1. Inference: Basis on these predictions, what are the business insights and recommendations. | **20-21** |
| **Problem 2:**   Logistic Regression and LDA |
| 1. **Introduction** |  |
| * 1. Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it? Perform Univariate and Bivariate Analysis. Do exploratory data analysis. | **22-28** |
| * 1. Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis). | **29-30** |
| * 1. Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized. | **30-34** |
| * 1. Inference: Basis on these predictions, what are the insights and recommendations. | **34-36** |

**Equation :**

1. precision is t tp / (tp + fp)
2. recall = tp / (tp + fn)
3. accuracy = tp+tn/(tp + fp + tn + fn)

where tp is the number of true positives,

fp the number of false positives

tn false negative

fn false negative

**Problem 1**: Linear Regression

You are hired by a company Gem Stones co ltd, which is a cubic zirconia manufacturer. You are provided with the dataset containing the prices and other attributes of almost 27,000 cubic zirconia (which is an inexpensive diamond alternative with many of the same qualities as a diamond). The company is earning different profits on different prize slots. You have to help the company in predicting the price for the stone on the bases of the details given in the dataset so it can distinguish between higher profitable stones and lower profitable stones so as to have better profit share. Also, provide them with the best 5 attributes that are most important.

**Data Dictionary:**

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Carat | Carat weight of the cubic zirconia. |
| Cut | Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal. |
| Color | Colour of the cubic zirconia.With D being the worst and J the best. |
| Clarity | Clarity refers to the absence of the Inclusions and Blemishes. (In order from Worst to Best in terms of avg price) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1 |
| Depth | The Height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter. |
| Table | The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter. |
| Price | the Price of the cubic zirconia. |
| X | Length of the cubic zirconia in mm. |
| Y | Width of the cubic zirconia in mm. |
| Z | Height of the cubic zirconia in mm. |

* 1. **Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA, duplicate values). Perform Univariate and Bivariate Analysis.**

**Solution:**

The sample data, and perform basic checks

**A screenshot of a computer

Description automatically generated with medium confidence**

**Table 1. sample data**

**inferences**-

1. the data set has two integer variables, six floating point variable remaining all are object variable
2. 26967 row and 11 columns in the given data
3. The given data set has only depth variable has the 697 null values
4. The given data has no duplicated values
5. The given dataset has the 11 columns
6. The range of the index is 26967 from 0 t0 26966
7. Memory usage is 2.3 MB
8. Data types is float(6),int(2) and object(3)

**Univariate analysis** -

To perform Univariate analysis on continuous variable, let statistic of the dataset

Table

Description automatically generated

**Table 2. summary of the dataset**

**Observation:**

* On the given data set the mean and median values does not have much difference.
* We can observe Min value of "x", "y", "z" are zero this indicates that they are faulty values. As we know dimensionless or 2-dimensional diamonds are not possible. So, we have filter out those as it clearly faulty data entries.
* There are three object data type 'cut', 'color' and 'clarity'.

**Checking for the skew**

carat 1.116481

depth -0.028618

table 0.765758

x 0.387986

y 3.850189

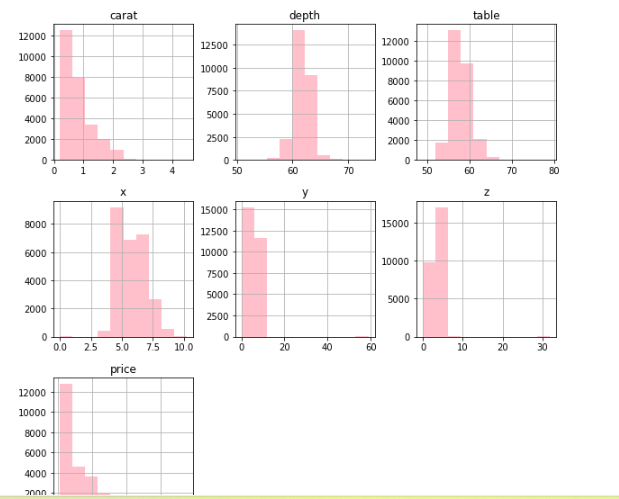
z 2.568257

price 1.618550

dtype: float64

* There is significant amount of outlier present in some variable.
* We can see that the distribution of some quantitative features like "carat" and the target feature "price" are heavily "right-skewed".

**Histogram**



**Fig 1. Histogram of the variance**

**Distrution and Box Plot**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

**Fig2 Distrution and Box plot for the variable**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

**Fig 3 box plot of the after treating outlier**

#### **Observation**:

* There are significant amount of outlier present in some variable,the features with datapoint that are far from the rest of dataset which will affect the outcome of our regression model. So we have treat the outliar. We can see that the distribution of some quantitative features like "carat" and the target feature "price" are heavily "right-skewed**"**

**Bivariant analysis**

Analysing the relationship among continuous variables by using pairplot and correlation heatmap

Graphical user interface, application

Description automatically generated with medium confidence

**Fig 4. Heatmap of variance**

**Diagram

Description automatically generated with medium confidence**

**Fig 5. Pairplot of the variance**

**Insights**

* It looks like most features do correlate with the price of Diamond.
* The notable exception is "depth" which has a negligble correlation (~1%). Observation on 'CUT': The Premium Cut on Diamonds are the most Expensive, followed by Very Good Cut.
  1. **Impute null values if present, also check for the values which are equal to zero. Do they have any meaning, or do we need to change them or drop them? Check for the possibility of combining the sub levels of a ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning.**

**Solution:**

In the given data set depth has 678 miss values other are not having missing variable

Treating the missing values with help of median

**Text

Description automatically generated**

After treating the missing values is

**A picture containing text

Description automatically generated**

**Checking for the values which are equal to zero.**

Number of rows with x == 0: 3

Number of rows with y == 0: 3

Number of rows with z == 0: 9

Number of rows with depth == 0: 0

Shape of the data (26967, 11)

**After removing the zero values is**

Number of rows with x == 0: 0

Number of rows with y == 0: 0

Number of rows with z == 0: 0

Number of rows with depth == 0: 0

Shape of the data (26958,11)

Observations

* here we are dropping the zeros values
* On the given data set the mean and median values does not have much difference.
* we can observe there are some amounts of 'Zero' value present on the data set on variable 'x', 'y' and 'z'
* this indicates that they are faulty values.
* As we know dimensionless or 2-dimensional diamonds are not possible. So, we need to filter out those as it clearly faulty data entries.
  1. **Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from stats model. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning.**

**Solutions:**

Encode the data (having string values)

lets check the unique counts of all Objects

**A picture containing table

Description automatically generated**

**Fig 6. unique counts of the categorical variables**

Converting objects to categorical codes.

**Text

Description automatically generated**

**Table3. info of the data**

**A screenshot of a computer

Description automatically generated with medium confidence**

**Table4. numerical data**

* After converting all categorical into the numeric data, drop the dependent variable ‘price’
* Let us break the X and y dataframes into training set and test set. For this we will use Sklearn package's data splitting function which is based on random function
* Split X and y into training and test set in 70:30 ratio
* Import Linear Regression machine learning library
* invoke the LinearRegression function and find the bestfit model on training data
* Let us explore the coefficients for each of the independent attributes

The coefficient for carat is 8867.168799330626

The coefficient for cut is 108.53627159308888

The coefficient for color is 273.59351802083734

The coefficient for clarity is 431.93681264714235

The coefficient for depth is -25.64785229004759

The coefficient for table is -17.95341502050796

The coefficient for x is -1507.0935504636386

The coefficient for y is 1191.8936211397026

The coefficient for z is -103.75734467688387

* The intercept for our model is -1067.729247117436
* R-square is the percentage of the response variable variation that is explained by a linear model.
* R-square = Explained variation / Total variation
* R-squared is always between 0 and 100%: 0% indicates that the model explains none of the variability of the response data around its mean.100% indicates that the model explains all the variability of the response data around its mean. In this regression model we can see the R-square value on Training and Test data respectively 0.9317074064221308 and 0.9298118073921455
* RMSE on Training data: 907.1312415459143
* RMSE on Testing data: 911.8447345328436

Scatterplot of predicted variable

Chart, scatter chart

Description automatically generated

Observation we can see that the is a linear plot, very strong corelation between the predicted y and actual y. But there are lots of spread. That indicated some kind noise present on the data set i.e Unexplained variances on the output.

* Linear regression Performance Metrics:
* intercept for the model: -3171.950447307667
* R square on training data: 0.9311935886926559
* R square on testing data: 0.931543712584074
* RMSE on Training data: 907.1312415459143
* RMSE on Testing data: 911.8447345328436
* As the training data & testing data score are almost inline, we can conclude this model is a Right-Fit Model.

**Linear Regression using stats models.**

**Table

Description automatically generated**

**Table 5. Stat model**

Parameters of the stat model

Intercept -1067.729247

carat 8867.168799

cut 108.536272

color 273.593518

clarity 431.936813

depth -25.647852

table -17.953415

x -1507.093550

y 1191.893621

z -103.757345

dtype: float64

**Inferential statistics**

Table

Description automatically generated

Table 6. Summary of the lml

underroot of mean\_sq\_error is standard deviation i.e. avg variance between predicted and actal traing and testing is 908.99 and 914

Chart, scatter chart

Description automatically generated

Fig 7 scatter plot

Using scalling values

The coefficient for carat is 1.1785409106006766

The coefficient for cut is 0.03472971127753299

The coefficient for color is 0.13416117479808767

The coefficient for clarity is 0.20358085654236638

The coefficient for depth is -0.010286781466753357

The coefficient for table is -0.01110628969808232

The coefficient for x is -0.48801383180998964

The coefficient for y is 0.38324720303191145

The coefficient for z is -0.02079855738800933

* The intercept for our model is -5.746644888976871e-16
* Model score - R2 or coeff of determinant- 0.9298844

Chart, scatter chart

Description automatically generated

Fig 8. Scaling scatter plot

carat ---> 110.39840803872023

cut ---> 9.735832884959047

color ---> 5.540857156266731

clarity ---> 5.4138910132055305

depth ---> 913.8577741168755

table ---> 740.1349247750014

x ---> 10304.893609112349

y ---> 9355.791313072286

z ---> 2111.99521642904

Insights

we can see that the from the linear plot, very strong corelation between the predicted y and actual y. But there are lots of spread. That indicates some kind noise present on the data set i.e Unexplained variances on the output.

Linear regression Performance Metrics:

intercept for the model: -1067.72

R square on training data: 0.931

R square on testing data: 0.929

RMSE on Training data: 908.99

RMSE on Testing data: 914.85

As the training data & testing data score are almost inline, we can conclude this model is a Right-Fit Model.

Impact of scaling:

Now we can observe by applying z score the intercept became -5.746644-16. Earlier it was -1067.72 the co-efficient has changed, the bias became nearly zero but the overall accuracy still same.

From statsmodels:

we can see R-squared:0.931 and Adj. R-squared: 0.931 are same. The overall P value is less than alpha.

* 1. **Inference: Basis on these predictions, what are the business insights and recommendations.**

Observation-1:

- Price' is the target variable while all others are the predictors.

- The data set contains 26967 row, 11 column.

- In the given data set there are 2 Integer type features,6 Float type features. 3 Object type features. Where 'price' is the target variable and all other are predictor variable.

- The first column is an index ("Unnamed: 0”) as this only serial no, we can remove it.

Observation-2:

- On the given data set the mean and median values does not have much difference.

- We can observe Min value of "x", "y", "z" is zero this indicates that they are faulty values. As we know dimensionless or 2-dimensional diamonds are not possible. So we have filter out those as it clearly faulty data entries.

- There are three object data type 'cut', 'color' and 'clarity'.

Observation-3:

- we can observe there are 697 missing values in the depth column. There are some duplicate row presents. (33 duplicate rows out of 26958). which is nearly 0.12 % of the total data. So, on this case we have dropped the duplicated row.

Observation-4:

- There is significant amount of outlier present in some variable, the features with datapoint that are far from the rest of dataset which will affect the outcome of our regression model. So, we have treated the outlier. We can see that the distribution of some quantitative features like "carat" and the target feature "price" are heavily "right-skewed".

Observation-5:

- It looks like most features do correlate with the price of Diamond. The notable exception is "depth" which has a negligible correlation (~1%). Observation on 'CUT': The Premium Cut on Diamonds are the most Expensive, followed by Very Good Cut.

we can see that the from the linear plot, very strong corelation between the predicted y and actual y. But there are lots of spread. That indicates some kind noise present on the data set i.e Unexplained variances on the output.

Linear regression Performance Metrics:

intercept for the model: -1067.72

R square on training data: 0.931

R square on testing data: 0.929

RMSE on Training data: 908.99

RMSE on Testing data: 914.85

As the training data & testing data score are almost inline, we can conclude this model is a Right-Fit Model.

Impact of scaling:

Now we can observe by applying z score the intercept became -5.746644-16. Earlier it was -1067.72 the co-efficient has changed, the bias became nearly zero but the overall accuracy still same.

From statsmodels:

we can see R-squared:0.931 and Adj. R-squared: 0.931 are same. The overall P value is less than alpha.

**Problem 2:** Logistic Regression and LDA

You are hired by a tour and travel agency which deals in selling holiday packages. You are provided details of 872 employees of a company. Among these employees, some opted for the package, and some didn't. You must help the company in predicting whether an employee will opt for the package or not based on the information given in the data set. Also, find out the important factors based on which the company will focus on particular employees to sell their packages.

**Data Dictionary:**

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Holiday\_Package | Opted for Holiday Package yes/no? |
| Salary | Employee salary |
| age | Age in years |
| edu | Years of formal education |
| no\_young\_children | The number of young children (younger than 7 years) |
| no\_older\_children | Number of older children |
| foreign | foreigner Yes/No |

**2.1 Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis.**

Table

Description automatically generated

Table 7 sample data of problem two

**inferences**-

* Holliday packages and foreign both are categorical data type, Salary, age, educ, no\_young\_children, no\_older\_children these are integer type data type
* 872 rows and 7 columns
* there are no null values in the given dataset
* The given data has no duplicated values
* there are 7 columns and 872 rows
* this data type has 47.8 KB memory location
* this data has 5 integer and 2 object data type
* the range of index is 0 to 871

**Univariate analysis** -

To perform Univariate analysis on continuous variable, let statistic of the dataset

Table

Description automatically generated

Table 7. summary of the data

**Checking for the Skew**

Holliday\_Package 0.161348

Salary 0.710966

age 0.146412

educ -0.095087

no\_young\_children 0.000000

no\_older\_children 0.872881

foreign 1.170906

dtype: float64

**Histogram**

**Chart, box and whisker chart

Description automatically generated**

**Fig 9. Histogram of problem two**

**Distibution plot**

**Chart, histogram

Description automatically generated**

**Fig10 dist plot problem two**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

**Fig11 box plot problem two**

**After treating the outliers**

|  |  |
| --- | --- |
|  |  |

**Fig 12 treating outlier of box plot**

**Bivariant analysis**

Analysing the relationship among continuous variables by using pairplot and correlation heatmap

Heatmap

**Graphical user interface, application

Description automatically generated**

**Fig 13. Heatmap of the problem two**

**Pairplot**

**Chart, calendar

Description automatically generated**

**Fig 14. Pairplot of the problem two**

 It looks like most features do correlate with the holiday package. The notable exception is "no\_younger\_childern " which has a negligble correlation (~1%). Observation on 'salary': is the high correlated

**2.2 Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis).**

**Checking the unique values of the categorical values**

**Text

Description automatically generated**

**After Encoding the values**

**Graphical user interface, text, application

Description automatically generated**

* **Train-Test Split and Copy all the predictor variables into X data frame. Taking 70:30 training and test set. Random number seeding for repeatability of the code**
* **Importing the library and Fit the model on original data i.e. before up sampling**

**Table

Description automatically generated**

**Linear Discriminant Analysis**

* **Prepare training and test data and do not scaling**
* **taking 70:30 training and test seta and Random number seeding for repeatability of the code**
* **import the LinearDiscriminantAnalysis**
* **Build LDA Model with the help of fit function**
* **Check Correlation values**

**Table

Description automatically generated**

**Table 9. correlation values**

**2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.**

**Logistic Regression**

* **Fit the Logistic Regression model**
* **pridict the x train model and pridict the x test model**
* **checking for probability of ytest**

|  | **0** | **1** |
| --- | --- | --- |
| **0** | 0.570131 | 0.429869 |
| **1** | 0.380445 | 0.619555 |
| **2** | 0.752807 | 0.247193 |
| **3** | 0.537490 | 0.462510 |
| **4** | 0.659950 | 0.340050 |

* **Accuracy - Train Data and test data is 0.6360 and 0.6526**
* **Confusion matrix on train data is The result is being telling us that there are 267+121 correct predictions and 156+66 wrong predictions**

array ([[267, 66],

[156, 121]], dtype=int64)

* **Confusion matrix on test data is The result is being telling us that there are 114+57 correct predictions and 24+67 wrong predictions**

**array ([[114, 24],**

[ 67, 57]], dtype=int64)

* **Classification report of train data**

Table

Description automatically generated

The observation

* precision is the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives.
* here precision is 0.63 is 0 and 1 has 0.65
* recall ratio tp / (tp + fn) where tp is the number of true positives and fn the number of false negatives.
* here recall is 0.80 is 0 and 1 has 0.44
* here f1 score has 0.71 and 0.52
* accuracy is the ratio tp+tn/(tp + fp + tn + fn) where tp is the number of true positives, fp the number of false positives, (tp + fp) where tn is the number of true negative and fn the number of false negative.
* accuracy is 0.64
* The F-beta score can be interpreted as a weighted harmonic mean of the precision and recall, where an F-beta score reaches its best value at 1 and worst score at 0.
* **Classification report of test data**

**Table

Description automatically generated**

Observations

* precision is the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives.
* here precision is 0.63 is 0 and 1 has 0.70
* recall ratio tp / (tp + fn) where tp is the number of true positives and fn the number of false negatives.
* here recall is 0.83 is 0 and 1 has 0.46
* here f1 score has 0.71 and 0.56
* accuracy is the ratio tp+tn/(tp + fp + tn + fn) where tp is the number of true positives, fp the number of false positives, (tp + fp) where tn is the number of true negative and fn the number of false negative.
* accuracy is 0.65
* The F-beta score can be interpreted as a weighted harmonic mean of the precision and recall, where an F-beta score reaches its best value at 1 and worst score at 0.

AUC and ROC curve of training data

Chart, line chart

Description automatically generated

**Fig 15 AUC AND ROC CURVE for training data**

AUC and roc curve of Testinf data

Chart, line chart

Description automatically generated

**Fig 17 AUC AND ROC CURVE for test data**

**Linear discriminant analysis(LDA)**

* **Build LDA Model with the help of fit function**
* **Predict the training and testing values**
* **classification report of train data**

**Table

Description automatically generated**

* **classification report of test data**

**Table

Description automatically generated**

* confusion matrix on training data

array([[270, 63],

[166, 111]], dtype=int64)

* confusion matrix on testing data

array([[116, 22],

[68, 56]], dtype=int64)

* **Accuracy of the testing data and testing data: 0.6262295081967213 and 0.6526717557251909**
* **Auc and Roc curve of training data is**

**Chart, line chart

Description automatically generated**

**Fig 17 AUC AND ROC CURVE for training data LDA**

* **AUC and ROC Curve of testing data is**

**Chart, line chart

Description automatically generated**

**Fig 18 AUC AND ROC CURVE for testing data LDA**

* 1. **Inference: Basis on these predictions, what are the insights and recommendations.**

Observation-1:

- The data set contains 872 row, 7 column.

-Holliday\_packages and foreign both are catogorical data type

-Salary,age,educ,no\_young\_children,no\_older\_children these are integer type data typ

-The first column is an index ("Unnamed: 0")as this only serial no, we can remove it.

Observation-2:

- On the given data set the the mean and median values does not have much differenc.

- There is no missing values and duplicated values in the given dataset

Observation-3:

- There are significant amount of outlier present in some variable,the features with datapoint that are far from the rest of dataset which will affect the outcome of our regression model. So we have treat the outliar. We can see that the distribution of some quantitative features like "educ" is low skewed "foreign" are heavily "right-skewed".

Observation-4 :

It looks like most features do correlate with the holiday package. The notable exception is "no\_younger\_childern " which has a negligble correlation (~1%). Observation on 'salary': is the high correlated

Observation 5: cosiderining accuracy of both model are good but LDA has better as comparing the Logestic regrasion

* precision is the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives.
* here precision is 0.63 is 0 and 1 has 0.65
* recall ratio tp / (tp + fn) where tp is the number of true positives and fn the number of false negatives.
* here recall is 0.80 is 0 and 1 has 0.44
* here f1 score has 0.71 and 0.52
* accuracy is the ratio tp+tn/(tp + fp + tn + fn) where tp is the number of true positives, fp the number of false positives, (tp + fp) where tn is the number of true negative and fn the number of false negative.
* accuracy is 0.64
* The F-beta score can be interpreted as a weighted harmonic mean of the precision and recall, where an F-beta score reaches its best value at 1 and worst score at 0.
* precision is the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives.
* here precision is 0.63 is 0 and 1 has 0.70
* recall ratio tp / (tp + fn) where tp is the number of true positives and fn the number of false negatives.
* here recall is 0.83 is 0 and 1 has 0.46
* here f1 score has 0.71 and 0.56
* accuracy is the ratio tp+tn/(tp + fp + tn + fn) where tp is the number of true positives, fp the number of false positives, (tp + fp) where tn is the number of true negative and fn the number of false negative.
* accuracy is 0.65
* The F-beta score can be interpreted as a weighted harmonic mean of the precision and recall, where an F-beta score reaches its best value at 1 and worst score at 0.