Credit Card Default Detection using Statistical Tools in Python

Divyanshu
21103104

Motivation

- After the demonetization in India, the number of credit card-holders are growing at the rate of 20% every year since 2016. So bank and financial institutions have to make sure that the person who are using credit card must be able payback the loan with in time.
- So rather than deciding manually about lending loans, machine learning and statistics can be used to make a model which can predict from given person's background details that if he/she will be able to payback the loan.
- Bharat pay, a unicorn startup is an example of a company who made a billion dollars by making use of machine learning and data analysis on data of small business owners.

Problem Statement

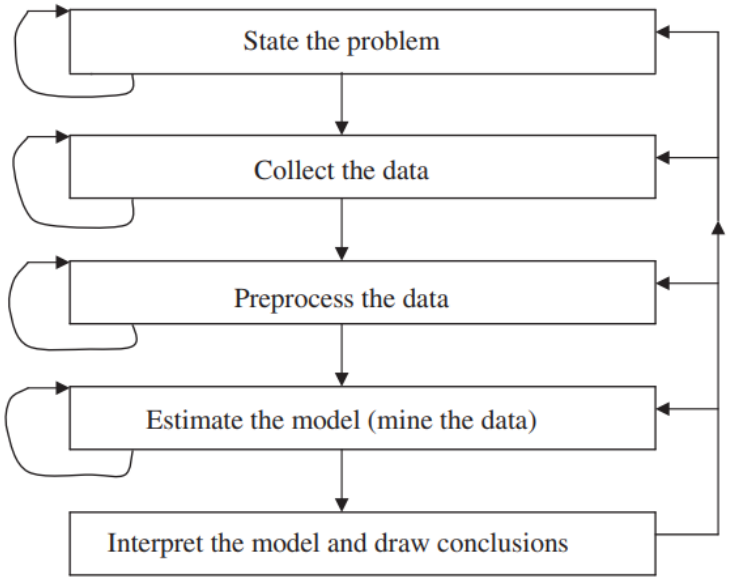
- **Credit Card Default Detection**
  Credit Card Default Detection by analysing various features using mathematical and statistical concepts and tools in Python.

Introduction of Data Mining

- **Definition**
  Data mining is the process of applying computational methods to large amounts of data in order to reveal new non-trivial and relevant information.
- **Two Goals of Data mining**
  1. Predictive Data-mining
  2. Descriptive Data-mining
- **Different fields in which data mining is used**
  1. Classification
  2. Regression
  3. Clustering
  4. Summarization
  5. Dependency modelling
  6. Change and deviation detection

Introduction of Data Mining Contd...

- **Fundamentals of Modeling**
  Fundamentals of Modeling is a concept of control theory engineering which can be used here. Two steps to make model is given below.
  1. Structure Identification
  2. Parameter Identification



Figure

Introduction of Data Mining Contd...

- **Steps of Data Mining**

1. Clear problem statement and hypothesis
2. Collect the data
3. Data pre-processing
   - 3.1 Outliers detection and removal
   - 3.2 Scaling, encoding and selecting features
4. Estimate the model
5. Interpret the model and make conclusions

Logistic Regression

- Logistic Regression is a supervised machine learning algorithm used in classification problems.

- **Examples**
  - To check whether tumour is malignant or benign
  - Hand-written digit recognition

Logistic Regression Contd...

- **Linear Regression and Logistic Regression**
  - **Hypothesis Function**

$$h_\theta(x) = \theta_0 + \theta_1 x = \theta^T x \tag{1}$$

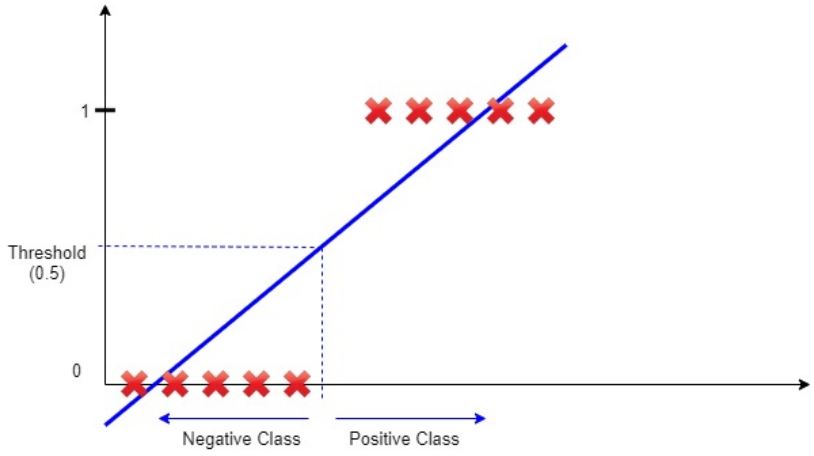
  - **Cost Function**

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2 = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)^2 \tag{2}$$

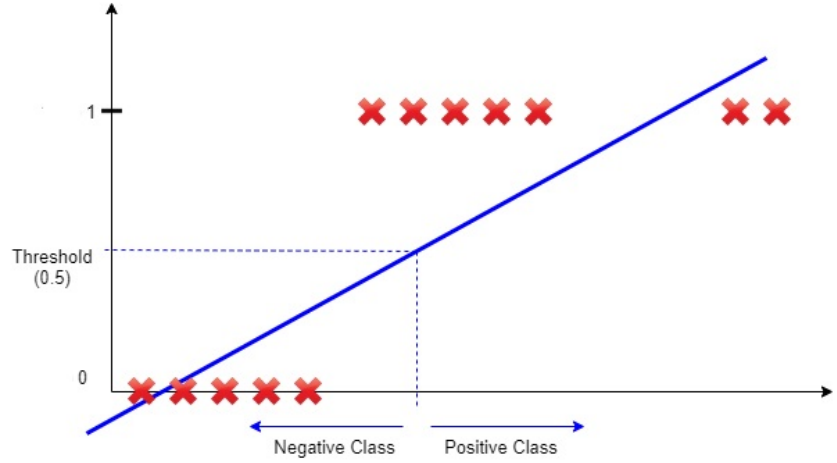
  - Problems
    - Problem with outliers
    - Gives value less than 0 and greater than 1

Logistic Regression Contd...

- **Linear Regression and Logistic Regression Contd...**



(a) Linear Regression



(b) Linear Regression with outliers

Figure: Linear regression

Logistic Regression Contd...

- **Logistic Regression**
  - **AIM** - $0 \leq h_\theta(x) \leq 1$

    $h_\theta(x) = g(\theta^T x)$, where $g(z) = 1/(1 + e^{-z})$ is called the sigmoid function or the logistic function. Therefore, $h_\theta(x) = 1/(1 + e^{-\theta^T x})$

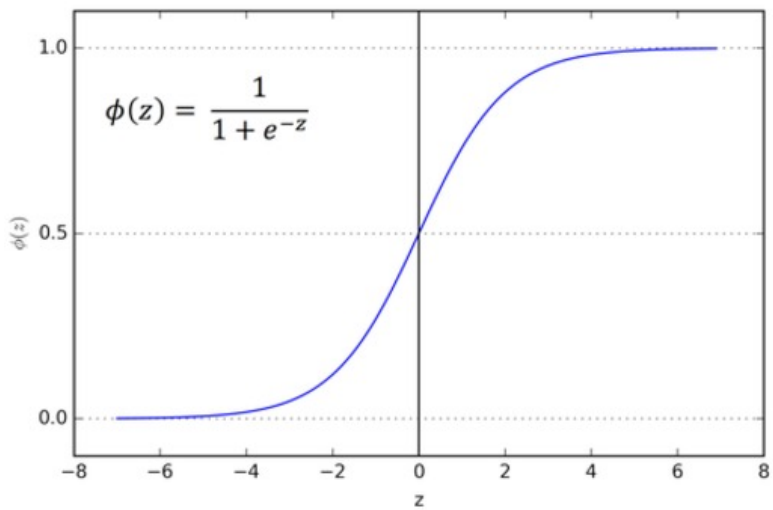


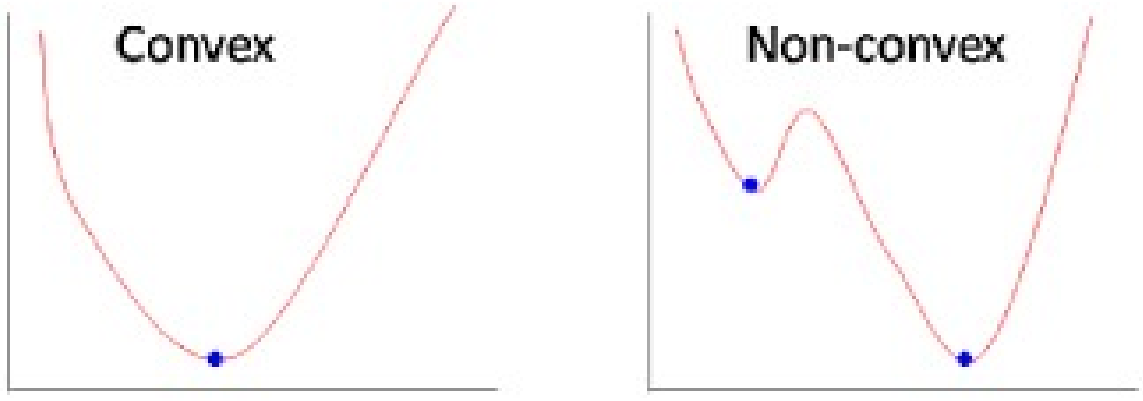
Figure: Sigmoid Function

  - **Cost Function**

$$Cost(h_\theta(x), y) = \frac{1}{2}[h_\Theta(x^{(i)}) - y^{(i)}]^2 \tag{3}$$

Logistic Regression Contd...

- **Logistic Regression Contd...**



Figure

- **Cost Function for Convex Curve**

$$Cost(h_\theta(x), y) = \left\{ \begin{array}{ll} -\log(h_\theta(x)), & \text{if } y = 1 \\ -\log(1 - h_\theta(x)), & \text{if } y = 0 \end{array} \right\} \quad (4)$$

Logistic Regression Contd...

- **Logistic Regression Contd...**

$$Cost(h_\theta(x), y) = -y \log(h_\theta(x)) - (1 - y) \log(1 - h_\theta(x)) \quad (5)$$

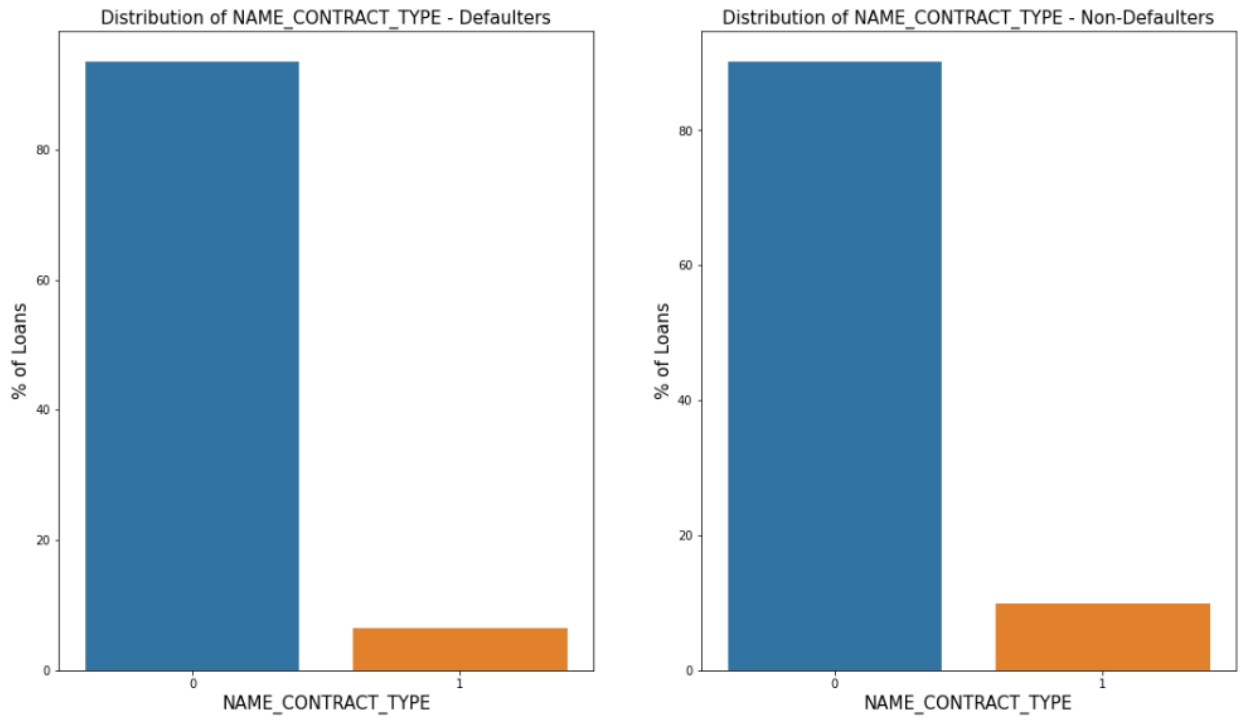
$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} Cost(h_\theta(x), y) \quad (6)$$

About Dataset

- Dataset is taken from kaggle
- This dataset contains information of 307511 people with 122 different features, so shape of the dataset is (307511, 122)
- Target Variable - Defaulters represented by 1 and Non defaulters represented by 0
- Presence of both quantitative and qualitative data
- Qualitative data has nominal and cardinal values

Relationship with target variable

- **Distribution of Cash loans and Revolving loans**

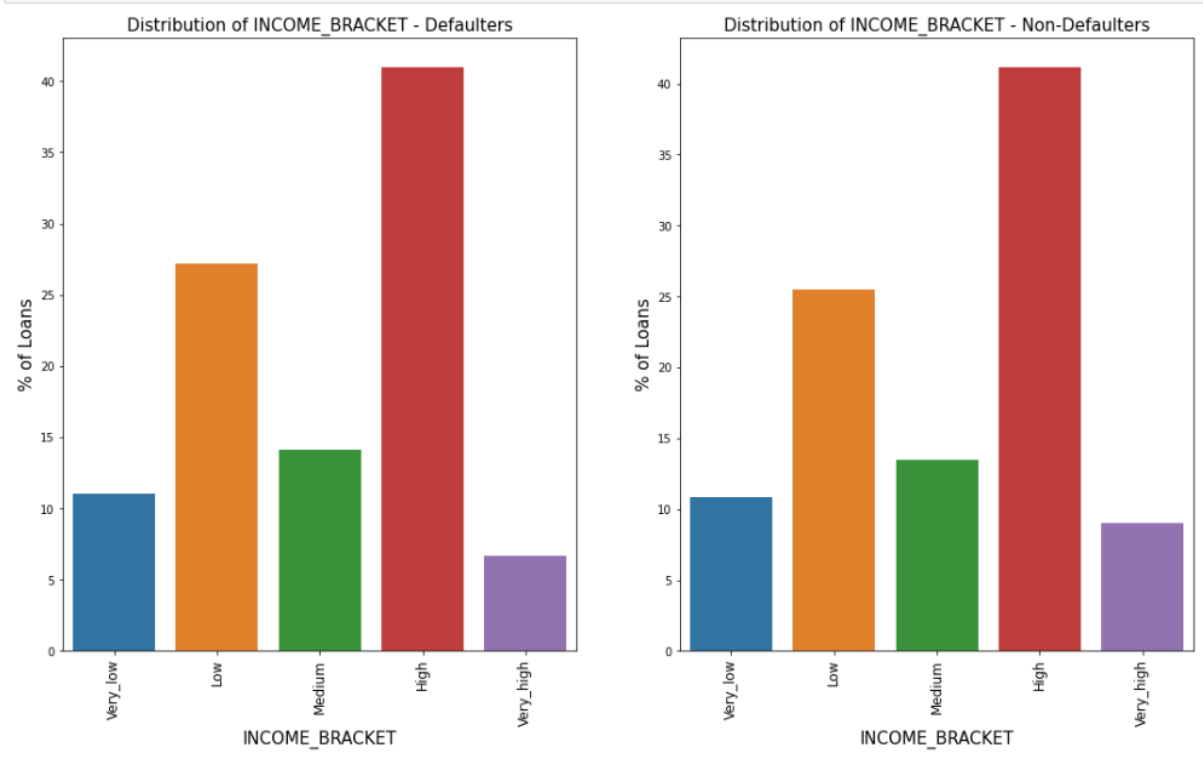


Figure

Number of cash loans are far more than revolving loans in both defaulters and non-defaulters

Relationship with target variable contd...

- **Distribution of different income brackets**



Figure

Here we can observe that number of people in low income and high income are high in both defaulters and non defaulters

Conclusion

- The model's accuracy, i.e., the number of predictions correctly made divided by the total number of predictions, is 91.92%.
- A Confusion Matrix in also constructed to get more insights about our model's performance.
- It is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known.
- The general idea is to count the number of times instances of class A are classified as class B.

Conclusion Contd...

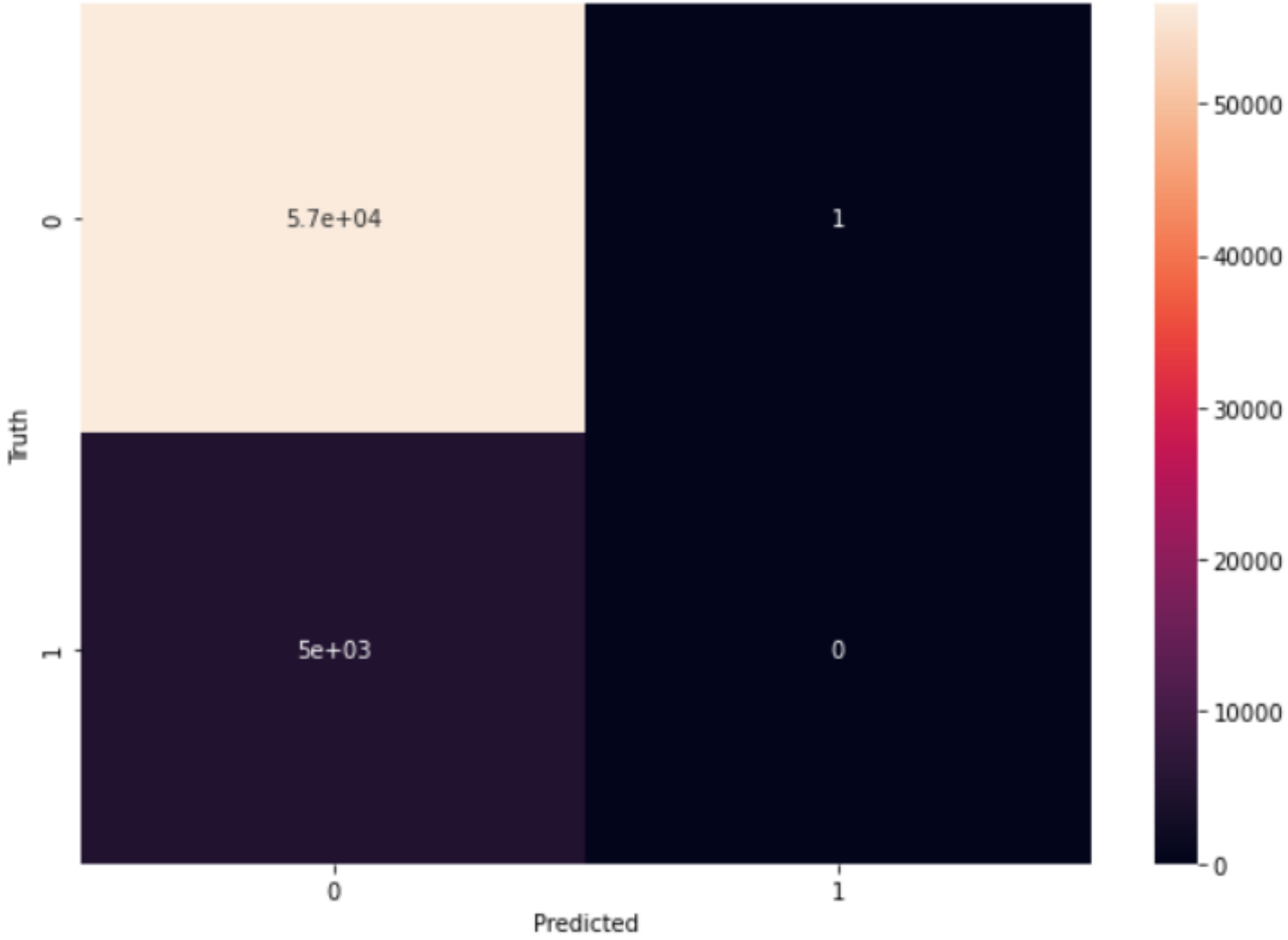


Figure: Confusion Matrix

Interpretation of Confusion Matrix

- For truth values 0, the model gave correct predictions as 0 itself a total of 56536 times (out of 61502 testing data). Therefore, the percentage of correct predictions is (56536/61502)*100 = 91.9254%.
- For truth value 1, the model predicted 0 4965 times, and the percentage is 8.0729%.
- Similarly, for truth value 0, it predicted 1 only once, the percentage being 0.00162%, and for truth value 1, it did not predict 1 even once, hence the percentage is 0.
- The fact that the model predicts over 90% of the test cases correctly is in our favor. Nevertheless, it can also be noticed that we have a case of overfitting, where the model has trained itself too well according to the training set and is having problems in generalizing new testing sets. This problem can be refined by using undersampling techniques and further looking into other measures, which are discussed in the next section.

Future Projections

- The over-fitting problem can be tackled by undersampling techniques.
- Other solutions - Cross-validation, training with more data, removing more irrelevant features, Regularization, Ensembling, etc.
- Other measures can also be calculated along with model accuracy, giving a better intuition of the model's performance. These include Confusion Matrix, Precision, Recall, and f1 score.

Future Projections Contd...

- From Confusion Matrix we get more metrics to evaluate the model,
  1. True Positive (TP): Target variable labeled as positive that are actually positive
  2. False Positive (FP): Target variable labeled as positive that are actually negative
  3. True Negative (TN): Target variable labeled as negative that are actually negative
  4. False Negative (FN): Target variable labeled as negative that are actually positive

Future Projections Contd...

- Precision is the ratio of correct positive predictions to the total predicted positives.

$$precision = \frac{TP}{(TP + FP)} \tag{7}$$

- Recall is the number of correct positive truths divided by the total number of elements that actually belong to the positive class (i.e., the sum of true positives and false negatives).

$$recall = \frac{TP}{(TP + FN)} \tag{8}$$

- It is often convenient to combine precision and recall into a single metric called the F1 score, which is the harmonic mean of precision and recall and thereby gives the model's overall performance.

$$F1\_score = 2 * \frac{(precision * recall)}{(precision + recall)} \tag{9}$$

Bibliography

[1] Data Mining: Concepts, Models, Methods, and Algorithms,Srivastava, Ashok N(2005)

[2] Encyclopedia of Systems Biology,Kallio, Aleksi and Tuimala, Jarno(2013)

[3] https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classifi

[4] https://www.geeksforgeeks.org/confusion-matrix-machine-learning/

[5] https://www.geeksforgeeks.org/understanding-logistic-regression/

[6] https://satishgunjal.com/binary_lr/

[7] https://www.analyticsvidhya.com/blog/2020/12/beginners-take-how-logistic-regression-is-related-to-l

Questions?

Thank you!