ISSN: 2366-1313

Logistic regression model using python

K. Swaroopa Rani, D. Jahnavi, B. Mounika, T. Venkata Krishna, T. Devi prasad, SK. Sharmila
Assistant Professor, Priyadarshini Institute of Technology & Science, AP, India.
Under graduate B. TECH CSE, Priyadarshini Institute of Technology & Science, AP, India.

Abstract:

Product reviews are valuable for upcoming buyers in helping them make decisions. To this end, different opinion mining techniques have been proposed, where judging a review sentence's orientation (e.g. positive or negative) is one of their key challenges. Recently, deep learning has emerged as an effective means for solving sentiment classification problems. A neural network intrinsically learns a useful representation automatically without human efforts. However, the success of deep learning highly relies on the availability of large-scale training data. I propose a novel deep learning framework for product review sentiment classification which employs prevalently available ratings as weak supervision signals. The framework consists of two steps: (1) learning a representation (an embedding space) which captures the general sentiment distribution of sentences through rating information; (2) adding a classification layer on top of the embedding layer and use labeled sentences for supervised fine-tuning. I explore two kinds of network structure for modeling review sentences, namely, convolutional feature extractors and long short-term memory. To evaluate the proposed framework, I construct a dataset containing 1.1M weakly labeled review sentences and 11,754 labeled review sentences from Amazon. Experimental results show the efficacy of the proposed framework and its superiority over baselines.

1 INTRODUCTION:

Crimes are the significant threat to the humankind. There are many crimes that happens regular interval of time. Perhaps it is increasing and spreading at a fast and vast rate. Crimes happen from small village, town to big cities. Crimes are of different type – robbery, murder, rape, assault, battery, false imprisonment, kidnapping, homicide. Since crimes are increasing there is a need to solve the cases in a much faster way. The crime activities have been increased at a faster rate and it is the responsibility of police department to control and reduce the crime activities. Crime prediction and criminal identification are the major problems to the police department as there are tremendous amount of crime data that exist. There is a need of technology through which the case solving could be faster.

The Federal Bureau of Investigation (FBI) defines a violent crime as an offense which involves force or threat [1]. The FBI's Uniform Crime Reporting (UCR) program categorizes these offenses into four categories: murder, forcible rape, robbery, and aggravated assault. The FBI UCR program defines each of the offenses as follows: (i) Murder - The willful (non-negligent) killing of one human being by another. The UCR does not include deaths caused by accident, suicide, negligence, justifiable homicides and attempts to murder or assaults to murder (which are scored as aggravated assaults), in this offense classification [2]. (ii) Forcible Rape - Rape is a sexual attack on a female against her will. Though attempts or assaults to commit rape by threat or force are considered crime under this category, statutory rape (without force) and other sex offenses are excluded [3]. (iii) Robbery - The taking or attempting to take anything of value from the care, custody, or control of a person or persons by force or

threat of force or violence and/or by putting the victim in fear [4]. (iv) Aggravated Assault - It is the unlawful attack conducted by one person upon another to inflict severe or aggravated bodily injury. The UCR program specifies that an aggravated assault usually involves the use of a weapon or other means to produce death or great bodily harm. Attempted aggravated assaults that involves the use of guns and other weapons are considered to belong to this category because if the assault were completed, it would have leads to serious personal injury. An offense that involves both aggravated assault and larceny-theft occurring together, the offense is considered to belong to the category of robbery [5]. Unfortunate type of crimes to have become common place in the society. Law enforcement officials have turned to data mining and machine learning to aid in the fight of crime prevention and law enforcement. In this research, we implemented the Linear Regression, Additive Regression, and Decision Stump algorithms using the same finite set of features, on the communities and crime un normalized dataset to conduct a comparative study between the violent crime patterns from this particular dataset and actual crime statistical data for the state of Mississippi that has been provided by neighborhoodscout.com [6]. The crime statistics used from this site is data that has been provided by the FBI and had been collected for the year 2013 [6]. Some of the statistical data that was provided by neighborhoodscout.com such as the population of Mississippi, population distribution by age, number of violent crimes committed, and the rate of those crimes per 100K people in the population are also features that have been incorporated into the test data to conduct analysis. enforcement officials have turned to data mining and machine learning to aid in the fight of crime prevention and law enforcement.

2 RELATEDWORK:

Crime forecasting refers to the basic process of predicting crimes before they occur. Tools are needed to predict a crime before it occurs. Currently, there are tools used by police to assist in specific tasks such as listening in on a suspect's phone call or using a body came to record some unusual illegal activity. Below we list some such tools to better understand where they might stand with additional technological assistance.

One good way of tracking phones is through the use of a stingray [35], which is a new frontier in police surveillance and can be used to pinpoint a cellphone location by mimicking cellphone towers and broadcasting the signals to trick cellphones within the vicinity to transmit their location and other information. An argument against the usage of stingrays in the United States is that it violates the fourth amendment. This technology is used in 23 states and in the district of Columbia. In ref. [36], the authors provide insight on how this is more than just a surveillance system, raising concerns about privacy violations. In addition, the Federal Communications Commission became involved and ultimately urged the manufacturer to meet two conditions in exchange for a grant: (1) "The marketing and sale of these devices shall be limited to federal, state, local public safety and law enforcement officials only" and (2) "State and local law enforcement agencies must advance coordinate with the FBI the acquisition and use of the equipment authorized under this authorization." Although its use is worthwhile, its implementation remains extremely controversial. A very popular method that has been in practice since the inception of surveillance is "the stakeout". A stakeout is the most frequently practiced surveillance technique among police officers and is used to gather information on all types of suspects. In ref. [37], the authors discuss the importance of a stakeout by stating that police officers witness an extensive range of events about which they are required to write a report. Such criminal acts are observed during stakeouts or patrols; observations of weapons, drugs, and other evidence during house searches; and descriptions of their own behavior and that of the suspect during arrest. Stakeouts are extremely useful, and are considered 100% reliable, with the police themselves observing the notable proceedings. However, are they actually 100% accurate? All officers are humans, and all humans are subject to fatigue. The major objective of a stakeout is to observe wrongful activities. Is there a tool that can substitute its use? We will discuss this point herein.

3.METHODOLOGY:

Data randomization.

The training examples given may not be in random order, which may produce misleading results. Therefore, we need to randomize the dataset first before dividing it into validation subset and training subset. The feature matrix and label matrix are combined into one extended matrix, and the extended matrix is shuffled. In this way, the dataset can be randomized efficiently and correctly. Randomization is done once at the very beginning of the system. Generally, logistic regression is well suited for describing and testing hypotheses about relationships between a categorical outcome variable and one or more categorical or continuous predictor variables. In the simplest case of linear regression for one continuous predictor X (a child's reading score on a standardized test) and one dichotomous outcome variable Y (the child being recommended for remedial reading classes), the plot of such data results in two parallel lines, each corresponding to a value of the dichotomous outcome.

Parameter tuning:

For simplicity, we use a single validation subset instead of using cross-validation. After randomizing the training data, the dataset is divided into training subset and validation subset. The number of examples in validation subset is predefined. The value of the coefficient β determines the direction of the relationship between X and the logit of Y. When β is greater than zero, larger (or smaller) X values are associated with larger (or smaller) logits of Y. Conversely, if β is less than zero, larger (or smaller) X values are associated with smaller (or larger) logits of Y. Within the framework of inferential statistics, the null hypothesis states that β equals zero, or there is no linear relationship in the population. Rejecting such a null hypothesis implies that a linear relationship exists between X and the logit of Y. If a predictor is binary, as in the Table 1 example, then the odds ratio is equal to e, the natural algorithm base, raised to the exponent of the slope β.

Validations of predicted probabilities. As we explained earlier, logistic regression predicts the logit of an event outcome from a set of predictors. Because the logit is the natural log of the odds (or probability/[1–probability]), it can be transformed back to the probability scale. The resultant predicted probabilities can then be revalidated with the actual outcome to determine if high probabilities are indeed associated with events

and low probabilities with events. The degree to which predicted probabilities agree with actual outcomes is expressed as either a measure of association or a classification table. There are four measures of association and one classification table that are provided by SAS.

Verification of the Binomial Assumption

As stated earlier, logistic regression has only one assumption: The binomial distribution is the assumed distribution for the conditional mean of the dichotomous outcome. This assumption implies that the same probability is maintained across the range of predictor values. Though none of the eight studies verified or tested this assumption, the binomial assumption is known to be robust as long as the sample is random; thus, observations are independent from each other. Samples used in the eight studies did not appear to be random, nor did they have inherent dependence among observations. Thus, the binomial assumption appeared to be robust underlying all logistic analyses conducted by these eight studies. During the last decade, logistic regression has been gaining popularity. The trend is evident in the JER and higher education journals. Such popularity can be attributed to researchers' easy access to sophisticated statistical software that performs comprehensive analyses of this technique. It is anticipated that the application of the logistic regression technique is likely to increase. This potential expanded usage demands that researchers, editors, and readers be coached in what to expect from an article that uses the logistic regression technique. What tables, charts, or figures should be included? What assumptions should be verified? And how comprehensive should the presentation of logistic regression results be? It is hoped that this article has answered these questions with an illustration of logistic regression applied to a data set and with guidelines and recommendations offered on a preferred pattern of application of logistic methods.

4 RESULTS:

ISSN: 2366-1313

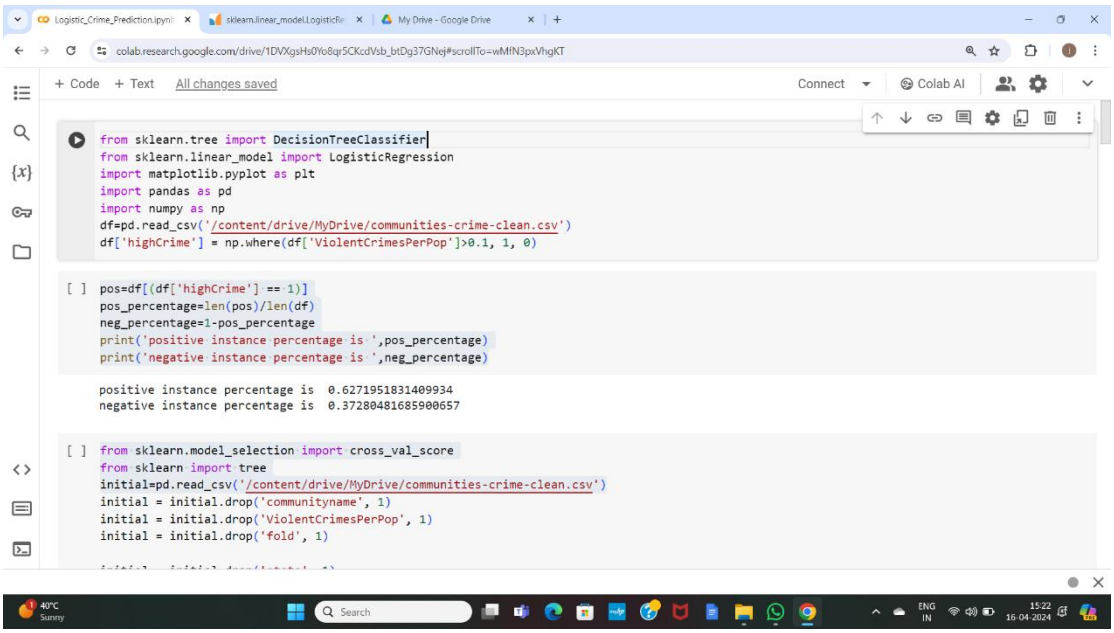

Fig 1: After connecting to internet package installed successfully.

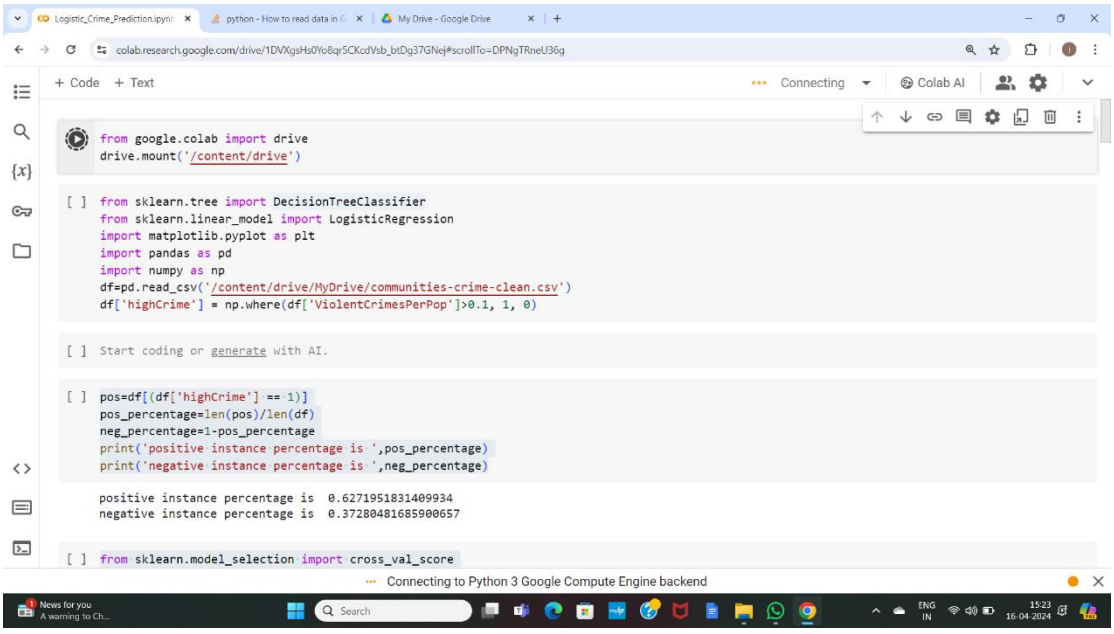

Fig 2:Installing another package successfully

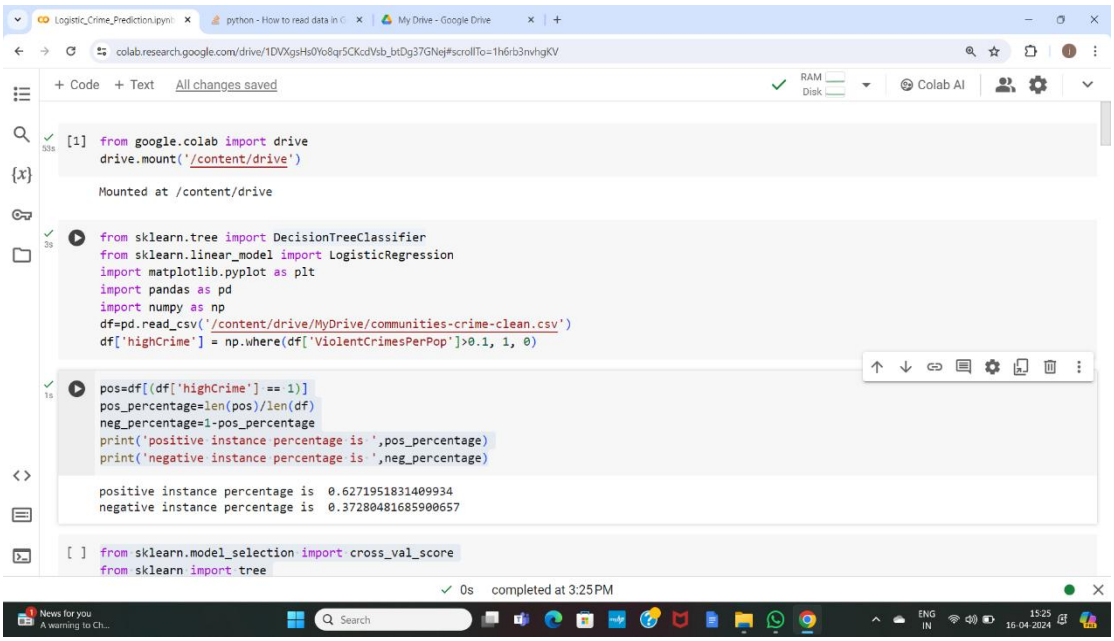

Fig 3: Here we get path not found error because it is not the correct path where the data is stored.

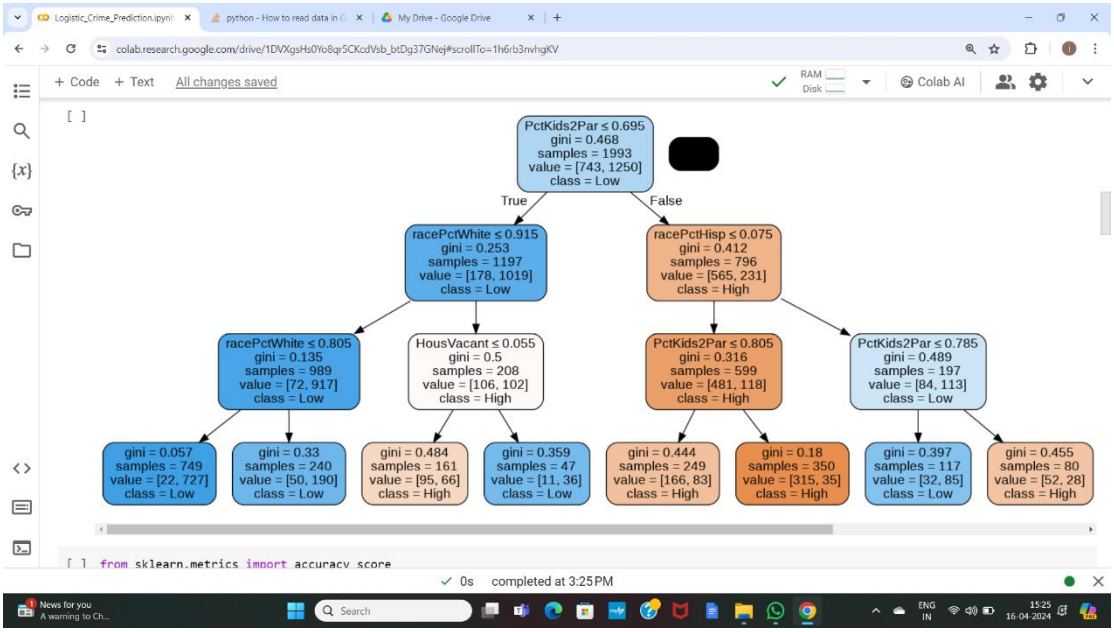

Fig 4Final result

5.CONCLUSION:

The objective of this study is to establish the microeconomic indicators which are able to predict the banking defect. The use of collected financial ratios from the Tunisian Banks balance sheets shapes our battery of indicators inspired by the CAMEL typology, from which we wanted to select the ratios that have a strong predictive power to construct a prevision model of bank defect from it. The use of a vector of ratios selected from advance by a stepwise regression, like a vector of explanatory

variables in our logistic model have provided us with satisfactory results with expected signs and significations. Likewise, the most pertinent ratios in the explanation of banking defect at the Tunisian banks are the decrease of banking profitability and the ability of banks to repay their debts which appear to be a high odd ratio.

6 REFERENCE:

[1] ATL under, Merve Büşra. (2014) “The Relationship Between Sociability and Household Debt.” Adam Academy Journal of Social Sciences/Adam Academic Social Biliml Dergisi4(2):27-58.

[2] André, Christophe. (2016). “Household debt in OECD countries: Stylish facts and policy issues.” OECD Economics Department Working Papers1277, OECD Publishing, Paris.

[3] Bagley, Steven C., White Halbert, Golomb Beatrice A. (2001) "Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain." Journal of Clinical Epidemiology, 54: 979–985.

[4] Bennouna, Ghita, and Mohamed Take out (2019) “Scoring in microfinance: credit risk management tool –Case of Morocco-.” Procedia Computer Science 148: 522-531.

[5] Boboli Piotr. (2020) “Developments in the household debt-to-GDP ratio across the OECD countries since global financial crisis.” Acta Sci. Pol. Economic 19(1): 5-12.

[6] Bodie, Zvi, and Robert C. Merton. (1998), Finance, Prentice Hall, Upper Saddle River, New Jersey, p. 4.

[7] Breuer, Wolfgang; Thorsten Hens; Astrid Juliane Salzmann, and Mei Wang. (2015) “On the determinants of household debt maturity choice.” Applied Economics 47(5): 449-465.

[8] Chien, Yi-Wen, and Sharon A. DeVaney. (2001) “The effects of credit attitude and socioeconomic factors on credit card and installment debt.” The Journal of Consumer Affairs 35(1): 162-179.

[9] Collins, J. Michael, Erik Hambr, and Carly Urban. (2020) “Exploring the rise of mortgage borrowing among older Americans.” Regional Science and Urban Economics 83: 1-23.

[10] Dobosz, Marek. (2004) Statistics Analiza Window. Academic Official Window EXIT, Warszawa.

[11] Ebrahimi, Zahra. (2020) "The impact of rising household debt among older Americans." EBRI Issue Brief 502: 1-22.

[12] Elvery, Joel A., and Mark E. Schweitzer. (2020) "Partially disaggregated household-level debt service ratios: construction, validation, and relationship to bankruptcy rates." Contemporary Economic Policy 38(1):166-187.

[13] Garriga, Carlos, Bryan Noeth, and Don Schlagenhauf. (2017) "Household Debt and the Great Recession." Review 99(2): 183-205.

[14] Haughwout, Andrew F., Dong Lee, Joelle Scally, Lauren Thomas, and Wilbert van der Klaw. (2019) "Trends in Household Debt and Credit." FRB of New York Staff Report 882.

[15] Hosmer, David W., Lemen show, Stanley, May, Susanne. (2008) Applied Survival Analysis: Regression Modeling of Time to Event Data. Wiley Blackwell.

[16] Hosmer, David. W., Leme show, Stanley. (2000) Applied Logistic Regression, New York: John Wiley & Sons.

[17] Jain, Hemlata, Ajay Khu tera, and Sumit Srivastava. (2020) "Churn prediction in telecommunication using logistic regression and logit boost." Procedia Computer Science 167: 101-112.

[18] Khan, Hafizah Hammad Ahmad, Hussin Abdullah, and Shamf Samsudin, (2016) "Modelling the Determinants of Malaysian Household Debt." International Journal of Economics and Financial Issues Eco journals 6(4): 1468-1473.

[19] Kim, Kyoung Tae, Melissa J. Wilmarth, and Robin Henager. (2017) Pover levels and debt indicators among low-income households before and after the Great Recession." Journal of Financial Counseling & Planning 28(2): 196-212.

[20] Kleinbaum, David, G., Klein, Mitchel. (2002) Logistic regression – a self-learning text. New York: Springer.

[21] Prasadu Peddi (2018), "A STUDY FOR BIG DATA USING DISSEMINATED FUZZY DECISION TREES", ISSN: 2366- 1313, Vol 3, issue 2, pp:46-57.

ISSN: 2366-1313