Title: Seattle Library Checkout Records

Electrical and Computer Engineering

ESOF-5011

Topics in Software Engineering: Big data machine learning models

Prof: - Dr. Emad Mohammad

Submitted by:

Names**:** Amarender Reddy Patel (0861386)

Sai Sumanth Kambala (0699712)

Anusha Kotti (0863993)

Submission Date**: - 2018/12/03**

Table of Content:

List of figures:

Abstract:

In this paper we are building a logistic regression model for the data to check weather a book will be checked out in a month. The data is obtained from the checkouts list of the books in the “Seattle Public Library(SPL)”. We are considering the checkouts for the past four years and we used “tableau software” to combine the data of four years and also to separate the data so as to save the time. We have explained the data with the tables and graphs and this data is clearly explained. We applied few data cleaning techniques to prepare the data for model building. Later we had built a logistic regression model for the data depending on the training data and later we tested the data that we have considered as the test data. Now in the final step we calculated the accuracy of the model by calculating the confusion matrix which is used to calculate the accuracy. From this we have examined the accuracy from the data by comparing the predicted data with the test data.

Keywords: Logistic regression, data preprocessing, data visualization, methodology, predictions.

1. Introduction:

Seattle Public Library(SPL) is a public library system serving Seattle, Washington. It is established in the year 1890. As of 2017, the central library contains about 1200K books. There are more than 26 branches located in different places. People at Seattle, visits to library and borrow books, read at library, search for new collection of books, etc., Hundreds of people comes every day and borrow books to home. Seattle has a situation, though people borrow books every day there will be shortage of books of edition. To avoid the shortage in library and not to disappoint the people who visits to library regularly, Seattle planned to predict which book will be checkout for next month. So, they started collecting the data from 2005 to till date. The data consists of book Title, Bibnumber, author name, time and date, etc., Using this CSV data, we are going to predict the future book will be checked out next month. Now the actual question starts, How to Solve? To predict this situation, we have different methods to do. The best way that I prefer to do is Logistic regression. It is a regression analysis to conduct when the dependent variable is binary. It helps me to predict the checkout data using month column as my input source. The methodology and the predicted ways can be explained in detailed in this project.

2. Related Works:

In this paper, we have a data based on the reading habit and depression tendency of a students. To build a psychological prediction, the study uses Linear regression and Logistic regression. The accuracy that was compared with different error criteria. The raw data includes depression score, sex, grades, borrowing frequency of technical book and different types of books. In this study, they use Scikit- learn in Python to construct linear and logistic regression [1]. Steps to follow are Load Standardized data and use dataset module to transform data into a format, loading the model, Training the model and later testing it. The count accuracy of predicting relative error at different depression score. The relative error can be identified using...

$$\varepsilon = \frac{|Yi - f(Xi)|}{Yi}$$

Y_i represents the actual values of the i^{th} item in the test data. $F(X_i)$ represent the predicted value of i^{th} term of the data. The relative error is larger the standard points they result the predicted result is wrong. The prediction accuracy of two regression is almost equal but the logistic regression has more accuracy. The possible error in this experiment is size of experiment and the frequency of the reading the book habit. The main disadvantage in this paper is they didn't identify the exact reason for why student get depressed while reading books. They don't have a sufficient data. They feel to have additional data which can help to find the exact accuracy of being depressed.

This paper studies the logistic regression to describe the number of literatures in a specific field. Analysing the shortcoming of improvements in logistic algorithm and proposed a new algorithm named DGA-logistic algorithm [2]. It is based on multiple objective genetic algorithm. This paper chooses Chinese digital Library Literature Dataset which were published recently. DGA logistic

Algorithm is a searching method encode the possible solutions into a population of individual which can be done in three steps like Selective reproduction, Crossing over, mutation. This algorithm takes two objective functions like minimizing the sum of squares or maximizing statistical significance and calculate the optimum solution. The main advantage of this method is Iteration process which helps to minimize the unwanted data. In this method the accuracy will be high. This paper mainly worked on DGA logistic Algorithm and detailed explanation how iterations process helps to minimize the unwanted data.

The aim of this paper is to examine and compare the methods used for improving the diagnostic accuracy of serum PSA in Turkey. The predictors used to detection of prostatic carcinoma were identified by logistic and Bayesian network. Bayesian network [3] is a directed graphical model that represent the joint probability distribution over a collection of random variables. Here two-third of dataset was randomly selected for establishing a predictive model and one third was utilized for testing. The random splitting was done using stratification principle to ensure that the proportions of positive and negative classes remains same in both testing and training the original dataset. This process is repeated n times to overcome sampling bias. This analysis was performed in WEKA machine learning software. Netica software were used for visualization of Bayesian network. And the predictive performance was evaluated on the test. They used Mann-Whitney U test to check the difference between the two models is significant. And the test result shows that logistic regression perform better than the Bayesian network. The disadvantage with Bayesian network is to perform n times which is more complex.

This paper they have provided a efficient method for predicting potential customers churn form imbalanced dataset on orange telecom and UCI. The main contribution of this paper is to apply a binary logistic regression model (LRM), which can be used in the problem of imbalanced data prediction. The parameters estimated from the data has a severe problem of imbalance. Therefore they took a stratified sampling method and improved traditional logistic regression model parameters estimated methods. In recent years, many types of research show that using the ROC curve and AUC has apparent advantages to evaluate the performance of the imbalance dataset classification. Using UCI and Orange to experiment on representative public data sets, with ROC curves [4] and AUC value as the evaluation index of experiments, comparing the experimental results show that the presented method for telecom customer churn prediction has the stable promotion effect.

In recent years online shopping has overgrown. To achieve better sales and provide marketing support to online owners, authors of this paper have studied and implement predictive modeling about the possibility of the customer to buy a tablet PC in the online shop. It proved that Logistic regression modeling could be used in predicting the possibility of the customer to buy a commodity and help to the online shop owner to decide. They effectively integrate the logistic regression modeling and data mining and develop the predictive modeling in the computer by Oracle and SAS [5]. Due to the time constraints and the lack of historical data, the logistic regression modeling mentioned in the paper has some limitations. Predictive modeling which supports for decision-making need to choose the best and most suitable explanatory variables; this requires continuously revised and improved in practice. In this paper, they have used the K-

S(Kolmogorov-Smirnov) test to verify the modeling ability between good and bad customers. If the modeling K-S value reaches 30%, the modeling is valid. If the K-S value is above 30%, modeling has higher discrimination. In this paper, K-S values are over 33% so that it can be used in practical work.

In many companies' financial indicators are still dominant in credit evaluation. However, it is not enough for credit rating on financial ratios. This paper designs a set of enterprise credit rating index system with traditional financial ratios and non-financial factors [6]. They have used the logical regression approach to create the company credits rating model. As the model has introduced the non-financial factors into the credit rating index system, the companies overall discrimination rate reached 95% which high.it proves that this model can predict results accurately.

This paper describes the pedestrian detection method using feature selection based on logistic regression analysis. Haar-like, HOG, and BHOG, HOG's derivative feature, are used. Stepwise forward selection, backward elimination, and LASSO methods [7] are used for logistic regression analysis. As for the results, the BHOG + Haar LRM had the best performance as its detection rate was about 95% and its false positive rate was about 10%. And, the processing time of this LRM is about 1.22ms, including feature extraction. The advantages of LRMPD are efficiency, which arose because of the exclusion of meaningless features and the ease of interpretation. However, high FPR is the drawback of this system. So, improvement of FPR and reduction of processing time are needed.

Shenzhen power supply bureau required a series of methods which can analyze the cause of wireless communication fault and predict wireless communication failure to solve the problem of signal fault in time and increase the online terminal rate. Authors of this paper have presented a model named GCFPM (Gradient descent to iterative Calculate optimal regression coefficient in power-grid communication Failures Prediction Model) which based on Logistic Regression Algorithm (LRA)[8], which effectively predict the possibility of communication failure, enhance the level of acquisition terminal operations, increase the terminals online-rate. Results and evaluations show that this model can perform well for predicting the power grid communication failures although it has several shortcomings. In the future, they choose other machine learning algorithms compared with LRA to improve the performance of their model.

3. Data:

3.1 Data Source:

Seattle public library started entering the checkout list from April 2005 and is regularly updated. The dataset consists of raw data which holds checkout time and drop few columns to reduce the size of a data. The data directory allows to decode the item type column. The data is updated every month, so we can forecast the data by downloading the new data. The data contains millions of rows by December 2017. The data was collected from the Kaggle site: https://www.kaggle.com/seattle-public-library/seattle-library-checkout-records. Hence the data contains more than 10000k rows, we considered data from 2014 to 2017 to shorten the risk, and the data that available is generated each year separately. So, to combine the data into a single .csv file we used a tableau to merge the data and splitting the checkout time columns into the multiple columns according to date, month, year and time. Tableau is a software used to handle 100 million+ rows. It helps us to analyses the dataset using graphs worksheets.

3.2 Data Visualization:

In this project our objective is to find weather a book that is selected will be checked out or not. For that we need a variable from the previous data to show a book is checked out or not. Here we have a set of columns in our data are “BibNumber, ItemBarcode, Itemtype, Collection, CallNumber, y”. The column “y” is constructed from the data given using a logic that a book is checked out in a month. In these columns the data are independent from each other and they are mostly identical. Now we can consider these columns from the data can be considered as independent variables and the column y as the dependent variable. Now depending on the variable from the data we need to consider a month, we have chosen a month and got the variable in logic format as shown in the data. As we know that the output variables must be in the format of (0,1) in the logistic regression model so we have selected an output as Y column.

The data we have considered has the following columns and as shown below.

```
data.columns
```

```
Index([u'BibNumber', u'ItemBarcode', u'ItemType', u'Collection', u'CallNumber',
       u'y'],
      dtype='object')
```

Fig1: Provides the column names

These table shows the column names that are present in the dataset that we have considered. In this data we considered seven columns out of which six of them are independent of each other and the column ‘y’ is dependent on each of them.

```
data.head()
```

	BibNumber	Call Number	item Type	ItemBarcode	Collection	y
0	2866447	282.092 D3308D 1997	cab	10077659265	acbk	0
1	2680408	B AL276F 2011	cab	10073115320	acbk	1
2	475976	B AL445	cab	10010117645	acbk	0
3	2450371	B Ad185A 2007	cab	10058490714	acbk	0
4	48815	B An434A	cab	10082769190	acbk	0

Fig2: Represents top 5 rows of the data.

This table shows the top 5 rows of the data that we have considered. From this we could say that the different columns that we are considering.

```
In [78]: data.dtypes

Out[78]: BibNumber       int64
         ItemBarcode     int64
         ItemType       object
         Collection     object
         CallNumber     object
         y              object
         dtype: object
```

Fig3: shows the layout of the object.

This table shows the layout of the object that is present for each column. That means it explains about the type of the data (weather it is an integer, float, object, etc.), size of the data (about the number of bytes), byte order and the shape.

```
data.shape
```

```
(25726620, 6)
```

Fig4: Shows the number of rows and columns present.

The above table shows the number of rows and columns that are present in the data that we have. As we can see from above that we have twenty-five lakh rows and six columns are present in the data.

```
: # Here we are trying to find the count the distinct values in each column
  data.apply(lambda x: len(x.unique()))
```

```
: BibNumber       495165
  Call Number     404121
  item Type          232
  ItemBarcode    2190472
  Collection          53
  y                    2
  dtype: int64
```

Fig5: Shows number of distinct values.

Table explains the number of distinct values in each column of the data and this shows the number of distinct values in the data that we had taken. From the above data we can see that the columns can be as independent variables and the last column as dependent variable.

```
: # to check the datatype and size of data
  data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25727877 entries, 0 to 25727876
Data columns (total 6 columns):
BibNumber      int64
Call Number    object
item Type      object
ItemBarcode    int64
Collection     object
y              object
dtypes: int64(2), object(4)
memory usage: 1.2+ GB
```

Fig6: Provides information about the datatype

Table explains about the information about the datatype that we are using and the size of the data that we are using. Here the size of the data that we use is more than 1.2 GB.

```
Text(0.5,1,'BibNumbervsValue_counts')
```

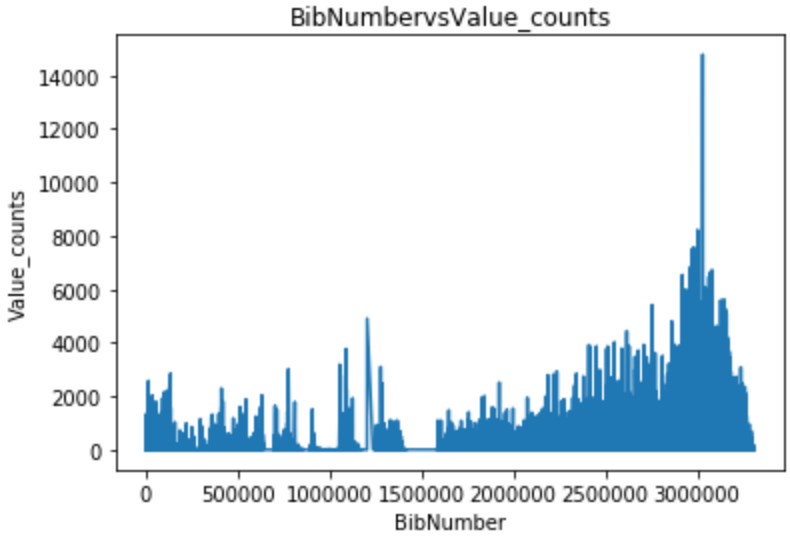

Fig7: Represents relationship between the BibNumber and its value counts

This graph builds a relationship between the BibNumber and its value counts. It shows the number of books as the value counts on y axis and the range of the Bibnumber on the y axis. From this table we could see the concentration of books.

```
Text(0.5,1,'ItemBarcodevsValue_counts')
```

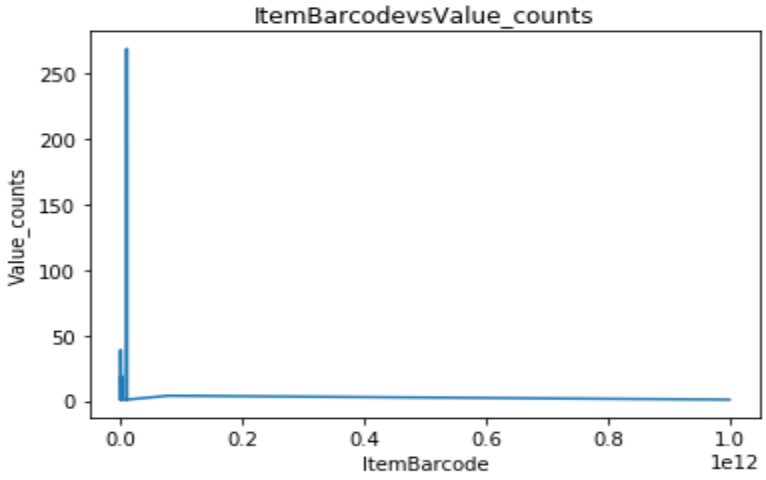

Fig8: Provides information between the ItemBarcode and its value counts.

This graph builds a relationship between the ItemBarcode and its value counts. It shows the number of books as the value counts on y axis and the range of the ItemBarcode on the y axis. From this table we could see the concentration of books.

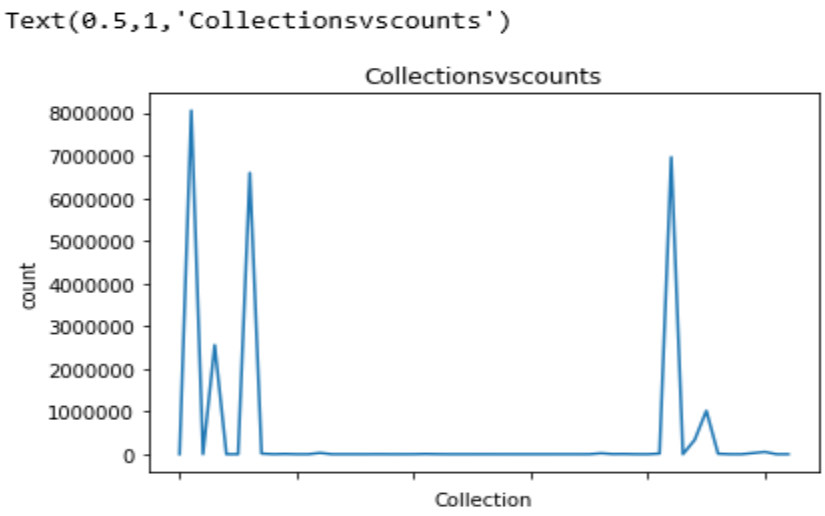

Fig9: Represents the relationship between the Collection and its counts.

This graph builds a relationship between the Collection and its counts. It shows the number of books as the value counts on y axis and the range of the Collection on the y axis. From this table we could see the concentration of books.

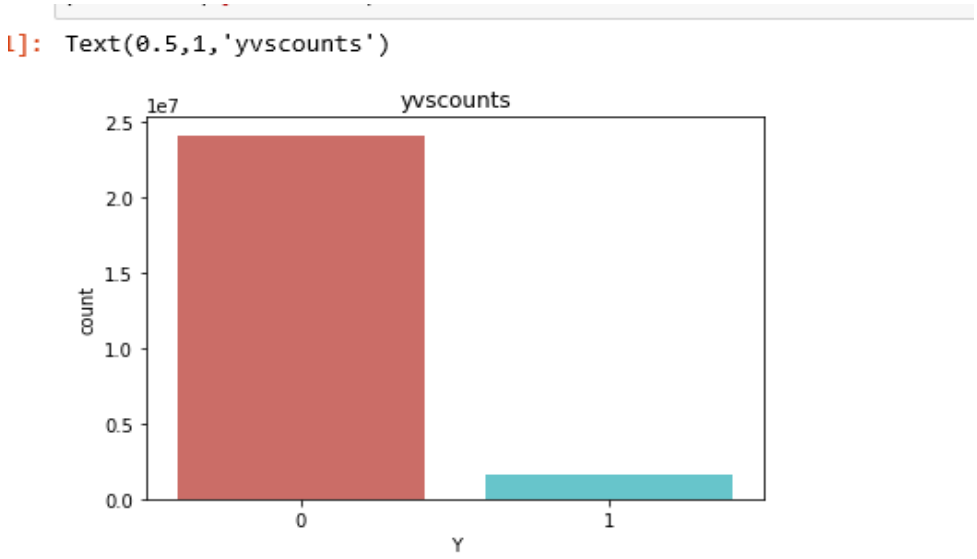

Fig10: Shows the count of 0's and 1's in data.

This barcode graph shows the y value counts in 0 and 1. Here the values of y are shown on the y axis and the binary on the x axis, from this we are showing the number of values y we have considered in the data are clearly explained.

4. Logistic Regression

Logistic regression is a statistical method used for analyzing a dataset in which the outcome is measured with a dichotomous. In general situations like this outcome are more common which may have one or more independent variables. Logistic regression is used to predict the likelihood of an outcome based on the input variables. The mathematically logistic model has a dependent variable with two possible values such as pass/fail, win/lose or alive/dead where the two values are givens as “0” and “1”.

Model description

Logistic regression is given by logistic function(y)

$$f(y) = \frac{e^y}{1+e^y} \quad for -\infty < y < \infty$$

The value of the logistic function f(y) varies from 0 to 1 as y increases.

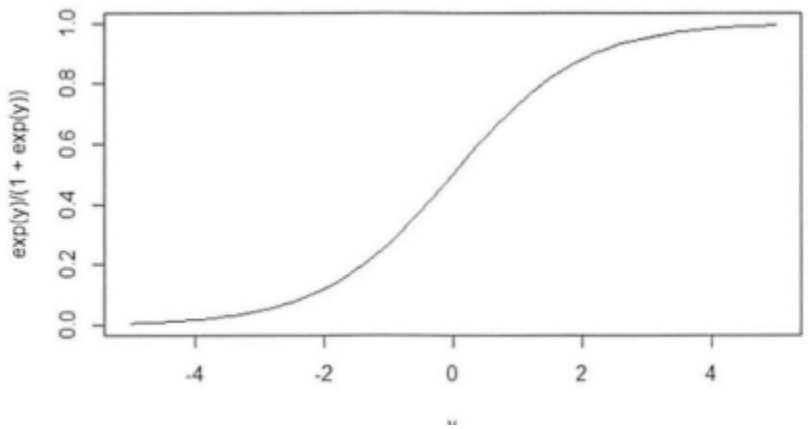

Fig11: Logistic function f(y) varies from 0 to1.

Consider the logistic regression model because the range of f(y) is (0,1), logistic function is the probability of a outcome occurring. The probability of the outcome increases with increases in y. in any proposed model, to predict the likelihood of an outcome, y should be the function of the input variables. Therefore

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots.. + \beta_{n-1} x_{n-1}$$

From the above equation, the probability of an event is given as

$$P(x_1, x_2, \ldots, x_{n-1}) = f(y) = \frac{e^y}{1+e^y} \quad for -\infty < y < \infty$$

The one main compared to difference here when compared to linear regression model is that values of y are not directly observed, it is only viewed regarding success or failure. Using p to donate f(y)

$$\ln(\frac{p}{1-p}) = y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{n-1} x_{n-1}$$

The above equation is known as log odds ratio, or the logit of p.in logistic model logit can be binary variable or a continuous variable. The corresponding probability can vary from 0 and 1. Maximum likelihood estimation is most commonly used to estimate the model parameters that maximize the chances of observing the given dataset. The main goal is to form an acceptable model which can provide the correlation between dependent and independent variables in best fit with the least variable. While measuring predictive variables to estimate the predicted variable over a probation time and using the obtained regression equation to calculate the predicted variable in the future are common practices. In both cases, when the nature of the correlation between variables X and Y is not fully known, it is essential to use data from the selected variables and define the correlation which could show the nature of the correlation to be used for prediction. The higher the number of variables to be included in a regression equation designed to explain the variance of the predicted variable, the less the error rate in the equation will be. However, problems and possible errors caused by workload for observation of each predictive variable and time limit for such observations would entail a decrease in the number of predictive variables. Therefore, the accuracy of predictions should be high as much as possible, and it is suggested to work with a reasonable number of predictive variables to lower systematic errors caused by data gathering using too many variables.

Logistic regression can be used in many fields, like machine learning, social sciences and in most medical fields. In the trauma and injury severity score used to predict mortality in injured patients using logistic regression. In other medical areas, it is used to develop to determine the likelihood of a patient's successful response to specific medical treatment. This technique can also be used in engineering for predicting the probability of failure of a given process, system or product. In marketing it is used to determine a wireless customers probability of switching carriers based on age, the number of family members in plan and social network contacts. It can be used in economics to predict the likelihood of a person willing to stay in the labor force, and the business application can be used to predict house owner is defaulting on a mortgage.

5. Data Preprocessing

The data that we obtained is from the four past years and the size of the data is about 1.2GB. As this data is huge first we have applied the preprocessing models for the sample data of 1 year and then we had these same for the whole data.

The first step in data preprocessing is finding out the null values as shown above, these can be removed, and the data is set for the next preprocessing steps. Here we have dropped all the column variables that we do not use these columns in our model. To reduce the complexity in the data we have reduced the columns by the following command.

```
# In this step we are removing the rows which are dependent and also to reduce the complexity of our data.
data.drop(['Month', 'Date', 'Hours', 'Min', 'AMPM', 'Year', 'XYZ-Split 8','XYZ','X','Table Name'], axis=1, inplace=True)
```

Fig12: Dropping of the columns.

The above table shows the columns that we are dropping to reduce the complexity in the data that we considering.

In the data we have few missing values, and this can be calculated by the table shown below.

```
]: # to replace not null value with 1 and null value with 0
   data.isnull().sum()

]: BibNumber         0
   Call Number    1257
   item Type         0
   ItemBarcode       0
   Collection        0
   y                 0
   dtype: int64
```

Fig13: Finding the null values in data.

From there we have removed the null values from the data by which we have got the data with no missing values. This gives us a complete set of data that is suitable for the test.

Now the data that we have is in the format of categorical data, but for building a logistic regression model we need to convert all the data to numerical and in the format of a matrix. This can be done by the following way.

```
|: #This helps to convert the categorical data to numerical data. this command selects the first column of the data.
   X[ : ,0] = label_encoder.fit_transform(X[ : ,0])

|: X

|: array([[333551, '282.092 D3308D 1997', 'cab', 10077659265, 'acbk'],
        [268190, 'B AL276F 2011', 'cab', 10073115320, 'acbk'],
        [39208, 'B AL445', 'cab', 10010117645, 'acbk'],
        ...,
        [58669, 'UNCAT', 'nafold', 10055943186, 'ucfold'],
        [58669, 'UNCAT', 'nafold', 10058834069, 'ucfold'],
        [90446, '746 D587M', 'canf', 10056735763, 'ucunkn']], dtype=object)
```

Fig14: Fixing labels to data.

The above table shows how the we are fixing labels to the data, in this the first column of the data is converted to numerical data. In the similar way we need to convert all the data to numerical data. This is shown in the table below.

```
X
```

```
array([[333551, 15456, 2, 981585, 1],
       [268190, 186204, 2, 727285, 1],
       [39208, 186224, 2, 41089, 1],
       ...,
       [58669, 384115, 151, 242565, 50],
       [58669, 384115, 151, 274045, 50],
       [90446, 111080, 45, 253257, 51]], dtype=object)
```

Fig15: Matrix format of the data.

From the above table now, the whole data is converted into matrix format with all the categorical data to the numerical. Now this data ready for the model building.

6. Model Building

As the data has been divided into training and testing for us to built a logistic regression model by the following way. These train data that we have has been used for building and the remaining data for testing.

```
logmodel = LogisticRegression()
```

```
logmodel.fit(X_train, y_train)
```

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
          penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
          verbose=0, warm_start=False)
```

```
# Considering the predicted value by using the input array values
predictions = logmodel.predict(X_test)
```

Fig16: Logistic regression model building.

This table explains how we have built a model using the training data that we had, we have separated the whole data into eighty percent for training and the rest we have used for testing the model.

```
#Grouping the input and output variables.
classification_report(y_test, predictions)
```

```
C:\Program Files (x86)\Microsoft Visual Studio\Shared\Anaconda3_64\lib\site-packages\sklearn\metrics\classification.py:1135: Un
definedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
```

```
'             precision    recall  f1-score   support\n\n          0       0.94      1.00      0.97   4827208\n          1
0.00      0.00      0.00    318116\n\navg / total       0.88      0.94      0.91   5145324\n'
```

Fig17: Classification report.

The *classification report ()* function from the above table show us the precision, recall, f1-score and support for each class of the data the we have considered for the y_test.

7. Results and Discussion

From this model we have obtained a model from which we could find a book weather be available in the coming month or not. This can be checked by calculating the accuracy of the model, this can be done by the test data that we had. In the previous section we had seen how we calculated classification matrix, now we will be seeing how to check the accuracy.

```
]: confusion_matrix(y_test, predictions)

]: array([[4827208,       0],
          [ 318116,       0]], dtype=int64)
```

Fig18: Confusion matrix.

From the table above, we have a confusion matrix which is helpful in calculating the performance of the model. From the confusion matrix we now be able to calculate the accuracy as follows

$$True\ positive = 4827208$$

$$True\ negative = 318116$$

$$\boldsymbol{Accuracy} = \frac{True\ positive}{TotalTrue\ positive + True\ negative}$$

$$= \frac{4827208}{5145324} = 93$$

```
: # the accuracy is calculates as the by adding True positive and falsenagative vales and divides by the total number of values.
  accuracy_score(y_test, predictions)

: 0.9381737670941616

: y1=y_test.iloc[:, ].values

: y1

: array(['0', '0', '1', ..., '0', '0', '0'], dtype=object)

: predictions

: array(['0', '0', '0', ..., '0', '0', '0'], dtype=object)
```

Fig19: Accuracy for the model.

As we have calculated above the accuracy is found to be 93.8 percent from the test data and the predictions that we have compared. We have shown the both the test data and the predictions in the array format.

The main advantage for our model accuracy is the data that we had, which helped us a lot in building a model of which we could calculate a good amount accuracy. The main problem is with the size of data that we used (which is more than 3GB). In this we have compared the usefulness of the data that we have.

8. Conclusion, Limitation, and Future work

From this paper we would be able to check weather a book from the library is checked out in a month. The model that we had built has an accuracy of about 93 percent, this can be achieved by having not much noise in the data and by the amount of data that we have considered. Here in the data preprocessing we have removed all the missing values where we lost few data. From this model we are unable to say the number of books that would be checked at once.

Future work: Seattle Public Library updates the checkouts data every month there contains a lot of information about the books that will be missing in the library. And, we can observe a new collection of books updated day to day. So, the prediction accuracy may be high when we have a huge data. In this paper, we predicted that the accuracy of checking out book in a month. Predicting the number of books of same kind will be checkout in a month will be the most challenging work in future. So that the Seattle library will be more careful about checking out list and can satisfy their customers.

References:

[1] Yujiao Hou, Jingjing Xu, Yixin Huang, et al “A Bigdata Application to Predict Depression in the University Based on the Reading Habit”, Third international conference on system and informatics(ICSAI), 2016.

[2] Xiao-Yong Liu, et al “An Improvement Logistic Model based on Multiple Objective Genetic Algorithm”. Eighth international conference on machine learning and cybernetics, Baoding, 12-15 July 2009

[3] Selen Bozkurt, Asli Uyar, et al “Comparison of Bayesian Network and Binary Logistic Regression Methods for Prediction of Prostate cancer” 4th International Conference on Biomedical Engineering and Informatics (BMEI) [2011].

[4] Li, Peng, et al. “Telecom Customer Churn Prediction Method Based on Cluster Stratified Sampling Logistic Regression.” International Conference on Software Intelligence Technologies

and Applications & International Conference on Frontiers of Internet of Things 2014, 2014, doi:10.1049/cp.2014.1576.

[5] Zhang, Weina, et al. “Research and Implementation of Predictive Modeling Based on Logistic Regression Modeling: About Possibility of Customer to Buy a Tablet PC.” 2012 Fifth International Joint Conference on Computational Sciences and Optimization, 2012, doi:10.1109/cso.2012.146

[6] Wei, Yong, et al. “The Listed Companys Credit Rating Based on Logistic Regression Model Add Non-Financial Factors.” 2010 Second International Conference on Modeling, Simulation, and Visualization Methods, 2010, doi:10.1109/wmsvm.2010.69.

[7] Kim, Jonghee, et al. “Optimal Feature Selection for Pedestrian Detection Based on Logistic Regression Analysis.” 2013 IEEE International Conference on Systems, Man, and Cybernetics, 2013, doi:10.1109/smc.2013.47.

[8] Liu, Tao, et al. “Prediction of Wireless Communication Failure in Grid Metering Automation System Based on Logistic Regression Model.” 2014 China International Conference on Electricity Distribution (CICED), 2014, doi:10.1109/ciced.2014.6991837.

[9] https://files.eric.ed.gov/fulltext/EJ919857.pdf

[10]
https://en.wikipedia.org/wiki/Logistic_regression#Logistic_regression_vs._other_approaches

https://www.medcalc.org/manual/logistic_regression.php

[11] Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data