Post Prediction On Facebook Page By Using Machine Learning Algorithm
Sonam, Surjeet Kumar

Abstract— In today’s era social network’s data is increasing and requiring highly managing services to handle large amount of data. We are seeing most of the people are mostly active on social network in which Facebook is the most popular. Professional users are especially using Facebook page for improving productivity and performance. Here, post prediction is made based on page category i.e., for a particular category of certain amount of posts. In order to predict all post and to find which type of post is getting the highest effect. We have analyzed with some machine learning techniques such as Logistic Regression and K Nearest Neighbor Classifier to make the prediction effective. This Paper predicts post for knowing popularity of facebook page on the basis of categories. Modeled the attributes that are available in data set such as User’s Interactions, Like and Share etc. Support Vector Regression Model is being used for graphical predicting value on data set and visualization of data is used by Heat Map and Extra Tree Classifier.

Index Terms— Heat Map, Extra Tree Classifier, Logistic Regression, K-Nearest Neighbor Classifier.

1 INTRODUCTION
Since Facebook is the most popular social network site that’s why we have chosen it as a research area and mining data from there. In section 2 we have illustrated about data characteristics and data set. We have described about algorithm in section 3 that is used in analyzing and predicting efficient value for post prediction such as how Logistic Regression, KNN algorithm, SVR Model, Extra Tree Classifier are more efficient techniques? All model’s graphical representation are useful for visualization of data and predict accurate value. Section 4 is the conclusion and section 5 is literature’s references that have been studied by us. Logistic Regression and KNN Classifiers have an efficient role in post prediction. These are supervised machine learning techniques required to split data for training, validating and testing classifiers. To find out best parameters of a classifier, training and validation are generally implemented with cross validation. Sigmoid function used in logistic regression that is bounded, differential analyzer, real function and defined for all real input values and has a positive imitative at each point. A sigmoid function and sigmoid curve refers to the similar object. SVR refers as Support Vector Regression whose basic idea is same as Support Vector Machine. SVR grants the presence of random data to provide an accomplished prediction model. A correlation matrix is a table expressing correlation coefficients between features. Each cell in the table displays the correlation between two categories. A correlation matrix is used to summarize and for briefing view of data, as an input into a more advanced analysis and as a diagnostic for progressed analysis.

2 DATASET AND DATA COLLECTION
The data we have taken that is available at https://archive.ics.uci.edu/ml/datasets/Facebook+metrics. Dataset to be needed as Import Data set contains the facebook.csv that expresses several design variables.

```
df = pd.read_csv('C:/Users/SONAM/Desktop/facebook.csv', index_col = 0)
```

- Sonam is currently pursuing Ph.D. program in Computer Application in VBS Purvanchal University, Jaunpur E-mail: sonamjha1980@gmail.com
- Surjeet Kumar is currently Assistant Professor in department of Computer Applications at VBS Purvanchal University, Jaunpur, India, PH-9415207124. E-mail: surjeet_k_yadav@yahoo.co.in

According to data set above mentioned, 19 attributes and 500 instances are available that is useful for calculating and predicting values. We are presenting Attributes and total values in table 1 that are used to predict accuracy.

<table>
<thead>
<tr>
<th>TABLE 1 DATASET CHARACTERISTICS</th>
<th>Total Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attributes</td>
<td></td>
</tr>
<tr>
<td>Page Total</td>
<td>61597088</td>
</tr>
<tr>
<td>type</td>
<td>610</td>
</tr>
<tr>
<td>category</td>
<td>940</td>
</tr>
<tr>
<td>post_Month</td>
<td>3519</td>
</tr>
<tr>
<td>post_Weekday</td>
<td>2075</td>
</tr>
<tr>
<td>post_Hour</td>
<td>3920</td>
</tr>
<tr>
<td>paid</td>
<td>139</td>
</tr>
<tr>
<td>lifetime_Post_Total_Reach</td>
<td>6951680</td>
</tr>
<tr>
<td>lifetime_Post_Total_Impressions</td>
<td>14792974</td>
</tr>
<tr>
<td>lifetime_Engaged_Users</td>
<td>460172</td>
</tr>
<tr>
<td>lifetime_Post Consumers</td>
<td>999386</td>
</tr>
<tr>
<td>lifetime_Post_Consumptions</td>
<td>707565</td>
</tr>
<tr>
<td>lifetime_Post_Impression_by_who_have_liked_page</td>
<td>8383188</td>
</tr>
<tr>
<td>lifetime_post_reach_by_who_like_page</td>
<td>3292744</td>
</tr>
<tr>
<td>lifetime_People_who_have_liked_page_and_engaged_with_post</td>
<td>304993</td>
</tr>
<tr>
<td>comment</td>
<td>3741</td>
</tr>
<tr>
<td>like</td>
<td>88795</td>
</tr>
<tr>
<td>share</td>
<td>13524</td>
</tr>
<tr>
<td>total_interactions</td>
<td>106060</td>
</tr>
</tbody>
</table>

3 EXPERIENTIAL DISCUSSION

3.1 Machine learning algorithm
The important role of Machine learning algorithm is in research, as it helps people discover new content and provide links to connect people. Machine learning and applied machine learning researchers develop machine learning algorithms that are useful in every work. It is of three types
- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
We have applied Logistic Regression and KNN Classifier algorithm which is a machine learning algorithm. This learning expresses in which we have need to taken input variable that is X and an output variable Y and algorithm is used to learn the depicting function from input to output.

\[ Y = f(X) \]

3.2 Logistic Regression
Logistic regression is a unique machine learning algorithm that
can be used to estimate discrete values such as 0 or 1, yes or no and true or false that is base on a given set of independent variables. Logical function used to predict the likelihood of an event, which will give output between 0 and 1. Since, logistic regression insert the data to a logit function i.e., logit regression.

3.3 How does Logistic Regression Works?
Logistic regression is a method that is use to analyze data set which has a dependent and one or more independent variables to predict results in a variable which is binary that means there will be two results. Logistic regression is a specific type of linear regression in which we predict results in absolute variables. It uses a logarithmic function to predict the likelihood of an event. We used sigmoid functions / curves to predict classification values. Threshold determines results.

Equation: \[ y = \beta_0 + \beta_1X_1 + \beta_2X_2 \ldots + \beta_nX_n \]

Sigmoid function: \[ p = \frac{1}{1 + e^{-y}} \]

3.4 K-Nearest Neighbors
KNN is machine learning algorithm that is simple, easy to know, versatile and uppermost. This is used in the several applications. Also used for both classification and regression problems and based on similar feature approach.

3.1.4 Work of KNN Algorithm
According to KNN, k is an arithmetical value of nearest neighbors. Value of neighbors is a basic determinant. If the value of neighbors is a basic determinant, the number of classes is 2, K is usually odd. If K=1, is called the nearest neighbor algorithm. That is the core case. Where P1 is the point label need to be predicted. At first, find out the point which is nearest to P1, after that search out the label of the nearest point assigned to P1 and then classify the points by the largest votes of its k neighbors. Every votes of an object for their class and the class have the maximum votes is used as a prediction. To find the nearest similarly at point, we can use space measurements to find the distance between points. Studies have shown that there is not accurate amount of neighbors suitable for every types of data sets. Each data set has its own requirements. If amount of neighbors is short, noise will have a greater impact on the results and the amount of neighbors will make the calculations large or extent. The study also shows that a few neighbors are the most flexible fit, which will have lower bias but higher variance and a largest number of neighbors will have straightforward decision boundaries that mean smaller variance but highest bias.

In general way, if the number of classes is even, the data scientist will choose it as odd. We can also check and check its performance by generating models on different values of k. According to Scikit-learn KNN classifier, Using Classifier construction, at first we have to define the data set and create or import dataset are needed for predicting value by using KNN algorithm.

3.5 How Does SVR Works?
SVR performs predicting values in showing steps:
- Gather a training set \( \{ X, Y \} \)
- Select a Kernel, its parameters
- Configure the correlation matrix, \( K \)
- Training the machine in approx way to find out contraction coefficients, \( = \{ i \} \)
- After that the use of coefficients to make estimator, \( f(X, x^*) = y^* \)

According to above equation, we are assessing the kernels of every point pairs in training set and add regulators to matrix. Algorithm is, \( K=y \) where \( y \) is the vector of values correspond to training set, \( K \) is correlation matrix then set would be not known that are needed to solve. Value is obtained by equation i.e. = K-1y

When parameters are known, we can design an estimator. In order to estimated unknown value \( y^* \), which is use to test point \( x^* \), we have requirement of inner product and the correlation matrix \( K \)

\[ y^* = K \]

We estimated elements of coefficient matrix by the equation:

Now SVR model is ready to predict unknown values.
SVR model takes value as an outlier. That is a cell containing 6.5 value is used to transform and inverse transform methods to convert the value of the feature scale to a normal value. In above SVR graph, magenta color is showing share and green color is showing like that predicted which values that are matching with value of data set.

3.6 Extra Tree Classifier for Feature Selection
The Extra Random Trees Classifier is an integrated learning technology that can summarize the results of multiple features collected in the form of graphical output in its classification results. At all test node, every tree would be provide with a random sample of k features according to feature set, and every decision tree has needed to be choose the best feature from this sample to split multiple features. Figure4 expresses Extra Tree Classifier that represents popular 10 features that is included on dataset. Each feature is shown in the form of color plot graph that is used for better visualization of features and useful in selection. Highest measurement shows which feature is important to categorize predicting value.

3.7 Visualizing Features by using Heat Map
Using paired graphs to draw relationships between many variables can quickly become overwhelmed visually. Therefore, it is often useful to calculate the covariance between variables. The covariance matrix can then be easily visualized as a heat map. The heat map is actually a pseudo-color image with labeled rows and columns that is a pseudo-color image.A heat map is a graphical representation of data, where data values are represented as colors. That is, it uses colors to communicate values to the reader. When we have a lot of data, it is a great tool to help audience look for areas that matter most.

We have used steps for correlation matrix using Heat map.

Create Correlation Matrix
corr_matrix = df.corr ()
We created correlation matrix with .corr. We have needed to dummy htype to calculate correlation
Needed have Set up Mask to Hide Upper Triangle
The upper triangle half of the correlation matrix is symmetrical to the lower triangle half. That’s why there is no need for heat map to show the entire matrix.
In this step, we have hidden the upper triangle
Let’s break the code down. np.zeros_like () returns an array of zeros with the same shape and type as the given array. By passing in the correlation matrix, we get an array of zeros. The dtype=np.bool parameter overrides the data type, so our array is an array of booleans.
np.triu_indices_from (mask) returns the indices for the upper triangle of the array. Now, we set the upper triangle to True.

3.8 Heatmap Using Seaborn
Using seaborn we have created Heat map and we passed in our correlation matrix from above step and the mask we created , along with custom parameters to make our heatmap look better.
Now, we have a mask that we can use to generate our Heatmap.

4. CONCLUSION
After observing, analyzing, visualizing and cleaning up data, a logistic regression model is constructed, divided it into train and test data, made predictions and finally evaluated them, one of the best properties of logistic regression is sigmoid function results the conditional probabilities of the prediction and the predicted value is 98%. Apart from this we also applied KNN supervised classification technique in which fit method is used to train the model and predict method. We have selected the value of K that is critical so it is to be needed for fit and test the model for different values and stored KNN’s testing accuracy in variable that is 98.33% then,
concluded that K-Nearest Neighbor Algorithm is performing well and giving the effective prediction.

**REFERENCES**


**ONLINE REFERENCES**

[1] https://www.academia.edu/Documents/in/Prediction
