Report

Data Science Project

Swarnim	Harsh Kumar Pal	Harshit Raj	Vipul Raj Jha	Nirmal Soni
2021567	2021047	2021051	2021435	2021xxx

Dataset

The dataset is related to websites URL which can be classified to phishing and non-phishing based on their attributes/features. It consists of 250000 rows and around 50 columns so together it makes around 10 million+ data entries. These are sufficient entries which are needed to work upon to feed up in ML algorithms and also to do some hypothesis testing. The last column in the dataset shows the genuineness of the website URL means it is the target column and contains values such as True or False. Other columns are either object or float64 type. Most of the columns with binary classes as entries are already converted into float64 type and contain 0 or 1 as their entries.

Problem Statement

We have to find the phishing URL on the basis of its attributes/features given in the dataset. We also have to establish some relationship between the phishing and non-phishing dataset groups.

Data Cleaning & Pre-Processing (Challenges with Dataset & It's Solution)

Missing Values & It's Removal

The dataset contains missing values both at random and in very high numbers in some columns like …. . At random, missing values are removed in two ways - first, dropping rows with more than 90% of columns having values as null, second, imputation of it using mean, median and mode. For columns with continuous values, impute with mean, for columns with discrete values, and highly skewed, impute with median and lastly for columns with discrete and non - skewed values, impute with mode. (Histogram, Bar Graphs plotted)

Outliers

The dataset is not very skewed but its individual columns are such that some are positively and others are negatively skewed so removing outliers is a challenge. If we do that the whole dataset will be lost. Hence, outlier removal is not done. (Box Plot plotted)

Removal of Uncorrelated Columns

Dataset has some columns which are purely string and don't have any relevance with the target column so dropping them off is the best possible option. Even some columns with very less correlation with target columns are dropped off to have only columns with high correlation with target. (Correlation Matrix plotted and Bar Graph between Columns & Targets)

Encoding for Categorical Columns
Any type of encoding is not required as categorical columns which mostly have binary categories are already encoded to 0 and 1 and simplify can be used to feed up in ML models and do hypothesis testing as binary digits works fine with these.

Also, as many binary classes containing columns are in the dataset so, while dropping columns (based on correlation with target), we prevent dropping of these with help of our dataset knowledge.

Hypothesis Test and their conclusion

We have done some hypothesis testing which included the F-test, Z-test and Chi-Square test.

Chi-Square Test
The chi-square test is done to know whether the categorical features are correlated to target columns or independent. First, we have sampled a dataset and then applied these tests on this sample in multiple numbers of iterations. The results of these iterations are stored in a list alongside the iterations. FInally, majority voting is used to get the final results of these tests. For validations, we take the whole dataset once and apply this test on it and save the result. If the result of this is equal to the result of majority voting then the test got validated and passed otherwise it fails in validation. In general, the 5% significance level is used to have the critical value of the chi-square test.

Z Test
It is done to find some type of relationship between the mean of two groups of website URL type, one is 0 (non-phishing) and another one is 1(phishing). We apply this test on a sampled dataset same as the chi-square test and then using majority voting of all results saved it. Then finally, try to compare means of group phishing and non-phishing on the whole dataset and if it comes the same as null hypothesis taken then test validates and passes otherwise test gets failed in validation. Null hypothesis here is basically that the mean of one group is lesser than the other. The Alternate hypothesis is that the mean of the other group is greater than the first one.

F Test
It is done to find some type of relationship between the variance of two groups of website URL type, one is 0 (non-phishing) and another one is 1(phishing). We apply this test on a sampled dataset same as the chi-square test and then using majority voting of all results saved it. Then finally, try to compare variance of group phishing and non-phishing on the whole dataset and if it comes the same as null hypothesis taken then test validates and passes otherwise test gets failed in validation. Null Hypothesis is here basically variance of one group is lesser than the other. The Alternative hypothesis is that the variance of the other group is greater than the first one.

Results
In most cases, validation held true for the dataset for f and t test. Only, have to drop one column with the help of chi-square test which shows its non-dependency with target columns.

Model Training & Performance Comparison (Scaled vs Unscaled)

Different ML models are trained to predict binary categories of target variables. Performance is compared both on the basis of train time, evaluation time and accuracy of different ML models. These models are SVM, Naive Bayes, Decision Tree, Logistic Regression and KNN.

Model Trained on Unscaled Dataset
Without scaling we have two types of ranking for ML models: first is accuracy wise and another one is running time (training + evaluation time) based. These rankings are Decision Tree, KNN, Logistic Regression, Naive Bayes, SVM (accuracy wise) and Naive Bayes, Decision Tree, Logistic Regression, KNN, SVM (running time based).

Model Trained on Scaled Dataset
Have used 3 different techniques to scale the dataset, first two manipulate the size of the dataset and 3rd one change the number of features.

SMOTE
We have applied SMOTE to balance the unbalanced dataset, it means to have the same number of rows with phishing and non-phishing groups by oversampling the lower one. It reduces variance as we are now not only predicting the target based on the majority class but on the basis of both classes.

Data Augmentation
This method adds random noise to the dataset to make it more generalized and reduce the variance while getting the model performance.

Applying these two also helps in increasing the f1-score indirectly balancing the trade-off between precision and recall or can say bias and variance.

SVD
This method is to break matrix of dataset into three matrix, orthonormal eigenvector of A*A^T, matrix of rank, same as dataset matrix (can take only r elements if needed) and transpose of orthonormal eigenvector of A^T*A. We only select r columns, r singular values and r rows for these 3 corresponding matrices to make a reduced dimension A which is our dataset matrix as others don't impact much on the result (can be shown when doing singular decomposition). It reduces the computation time for running the algorithm and also may increase accuracy in some cases as in our case, the performance of the model (here it refers to accuracy) increases somewhat, even its precision, recall and f1-score improves a bit.

Conclusion

We can at last conclude that we have found a ML prediction model to predict the phishing URL which has multiple applications and prevents users from frauds. This ML model (Decision Tree) works best for this dataset in terms of accuracy which is achieved to get the target prediction. Even, a relationship between the mean and variance of phishing and non-phishing groups are established using different hypothesis tests like f-test and z-test. Statistical tests like chi-square test helps in dropping categorical columns from dataset uncorrelated with target columns. Also, different plots help us to find patterns in the dataset like outliers, skewness, correlation between features and target columns. Even doing some other EDA helps us to know missing values (necessary to remove before training ML models and hypothesis testing) & statistical description about dataset to know where and whether to apply any encoding for categorical columns or not.

Future Work

At the same time, there will be a need to integrate various external datasets, which will enrich the dataset by embedding into it other aspects of achieving improvement in model predictive performance. At this point, efforts should be equally directed to the scalability and adaptability of the designed model. Deployment into different industrial environments will involve testing across larger, real-time scenarios concerning its performance under dynamic operational conditions. The optimization will keep the model relevant, robust, and able to solve emerging challenges, thus paving the way for its application in real and industrial contexts.

Code: https://github.com/SwarnimIIITD/DataScience-Project/blob/main/DSC_project%20(3).py
Dataset: https://archive.ics.uci.edu/dataset/967/phiusiil+phishing+url+dataset