International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

ELSEVIER
SSRN

IJIEMR Transactions, online available on 25th Jun 2022. Link
:http://www.ijiemr.org/downloads.php?vol=Volume-11&issue= Spl Issue 05

DOI: 10.48047/IJIEMR/V11/SPL ISSUE 05/18

Title **BUILDING A PREPROCESSOR CLI AND CRIME RATE PREDICTION USING TIMESERIES ANALYSIS**

Volume 11, SPL ISSUE 05, Pages: 117-123

Paper Authors
Mrs. T Padmaja, Pichika Lakshmi Likhitha, Tadiboina Brahnavi, Vunnamatla Daniel Paul, Veerla Ram Prasad

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

BUILDING A PREPROCESSOR CLI AND CRIME RATE PREDICTION USING TIMESERIES ANALYSIS

Mrs. T Padmaja[1], Pichika Lakshmi Likhitha[2], Tadiboina Brahnavi[3], Vunnamatla Daniel Paul[4], Veerla Ram Prasad[5]
[1]Assistant Professor, Dept. of IoT, [2]18ME1A0580, [3]18ME1A05B3, [4]18ME1A05B9, [5]18ME1A05C0
Ramachandra College of Engineering, A.P., India
padmajat77@gmail.com, likithapichika@gmail.com, tbrahnavi@gmail.com, mrdpaul015@gmail.com, ramprasadveerla@gmail.com

Abstract
Preprocessing data before feeding it into the model is crucial because the quality of data and the relevant information that can be derived from it has a direct impact on the model's capacity to learn. The main aim of this paper is to provide a basic CLI tool that will allow anyone to apply a variety of machine learning algorithms while saving time. Data preprocessing is a data mining technique that involves turning raw data into a comprehensible format. Data from the real world is typically insufficient, inconsistent, and/or missing in specific behaviors or trends, and it is likely to contain several inaccuracies. Pre- processing data is a tried and tested way to solve such issues.

Since the beginning of the 20th century, the Chicago Police Department's Bureau of Records has kept track of crime in the city. The total crime rate in the city, particularly the violent crime rate, is higher than the national average. The Prophet model has been used to predict the crime rate in Chicago. Facebook's Core Data Science team published Prophet as open-source software. It's a method for forecasting time series data that uses an additive model to match non-linear trends to yearly, weekly, and daily seasonality, as well as holiday effects. Prophet works best with time series with substantial seasonal effects and historical data from several seasons.

Introduction

Data pre-processing has generally been seen to be the most efficient method of making data more understandable to machines. Data is frequently incomplete in the actual world: it lacks attribute values, certain relevant qualities are absent, or it merely contains aggregate data. When data has errors or outliers, it becomes significantly noisier and inconsistent, similar to when there are inconsistencies in codes or names. This is why it is very important to preprocess the data before injecting it into the model.

Machine learning is the trendiest topic right now. Machine learning is a branch of Artificial Intelligence that focuses on teaching computers how to learn without being programmed for specific tasks. Indeed, one of the basic concepts of Machine Learning is that it is feasible to build algorithms that learn from and predict data. Everyone wants to jump on board with machine learning and implement it in their enterprises. This complex process revolves around data as it is the core of the data.

The effectiveness of the machine learning technology is determined by the quality of the data. Sophisticated algorithms will not compensate for poor data. Before being fit for use, data must go through various processing steps.

Machine Learning can be found everywhere including the spam filter that flags messages in email, the recommendation engine Netflix uses to propose content, and the self-driving cars being

developed by Google and other businesses.

Supervised learning, unsupervised learning, reinforcement learning, are types of Machine Learning. During this paper, the main goal is to write Python scripts to create a pre-processed dataset for supervised learning. The process of supervised learning involves mapping input data (independent variables) to known targets (dependent variables) provided by humans. In this paper, time series analysis has been used and one of the machine learning algorithms called the Prophet model to predict the crime rate in Chicago, where the total crime rate, specifically the violent crime rate, is higher than the US average.

The Facebook prophet is a time series forecasting model that is based on an additive regression model. It works well with data that has substantial seasonal influences as well as data from a variety of historical events. The Prophet is also robust to outliers, missing data, and abrupt changes in time series.

Prophet has two main advantages over other forecasting models: it is much easier to produce a fair, accurate forecast with Prophet, and Prophet forecasts are customizable in intuitive ways for non-experts.

Related Work

The command line is considered a relic of computing's past. The command line, on the other hand, is essential as a developer tool. Command Line Interfaces are frequently used to manage local processes and source control (CLIs). Another type of CLI has gained popularity in recent years, and it goes beyond the local system. For developers, the command line has evolved into a powerful tool for interacting with cloud services. CLIs are available for cloud computing services, continuous integration products, and some APIs. Pre-processors, as the name implies, are programs that process our source code before compilation. Between writing a program and executing it, there are several steps to take.

Learning algorithms have a preference for specific types of data, on which they excel. They're also known for making reckless predictions based on unscaled or unstandardized data.

Algorithms like XGBoost, in particular, require dummy encoded data, but decision trees appear to be unconcerned. Pre-processing, in simple terms, is the alteration of data prior to feeding it to the algorithm. The scikit-learn package in Python comes with pre-built functionality called sklearn. Preprocessing.

There are various data pre-processing steps such as:
importing the libraries → importing the data-set → checking out the missing values →encoding the categorical values → splitting the data → feature scaling.

Previously, the category variables were encoded using one of the encoding techniques known as Label Encoding. Each category is given a value ranging from 1 to N (where N is the number of categories for the feature). One major flaw with this approach is that there is no relationship or order between these classes, though the algorithm may be considered such. Various techniques such as Robust scaling and Maximum Absolute Scaling are used for feature scaling pre-processing steps.

Problem Statement

A. Existing System and its disadvantages

The performance of a machine learning model is determined by how different types of variables are processed and input into the model, as well as the model and hyperparameters. Because most machine learning models only accept numerical variables, preprocessing categorical variables becomes important. Categorical variables are finite in number and are commonly expressed as

'strings' or 'categories'.

There are two kinds of categorical data:

1. Ordinal Data: The categories have an inherent order.
2. Nominal Data: The categories do not have an inherent order.

Ordinal data should keep the information about the order in which the categories are presented while encoding. When encoding nominal data, the existence or absence of a feature must be considered. There is no notion of order in such a situation. In order for the model to understand and extract useful information, categorical variables should be converted to integers. Encoding the Categorical variable is one of the preprocessing steps in data preprocessing. Label encoding was previously used as an encoding technique. Because the label encoding process is simple and considers order when encoding, hence it can be used to encode ordinal data. The sequence should also be reflected in the label encoding. However,

Label Encoding has a disadvantage in that it considers column hierarchy, which might be misleading to nominal features present in the dataset.

Feature scaling is also another type of data preprocessing step. This step was previously done using Robust Scaling and Maximum Absolute Scaling techniques. The main disadvantage of Robust Scaling is that it ignores the median and just concentrates on the areas with the most data. The existence of huge outliers is one of the disadvantages of Maximum Absolute Scalar. In time series analysis, the present system for predicting Chicago crime rates uses the Multi-Season Holt- Winter model. However, this model has a
2.11 % error rate. Even so, because this model's accuracy rate is 91 %, it can be thoroughly optimized.

B Proposed System and its advantages

During this paper, the objective is to propose a system that overcomes the shortcomings of the existing system that were previously discussed. For the purpose of encoding the categorical variable, in this paper, t he technique of one-hot encoding was applied. One-hot encoding is the process of converting categorical input variables into machine and deep learning algorithms, which improves model predictions and classification accuracy. Normalization and standardization were used for feature scaling. Normalization is a scaling technique that shifts and rescales values and to make the values range between 0 and 1. The main advantage of normalization is, it drastically improves the model accuracy.

Another scaling technique is standardization, which center values around the mean with a unit standard deviation. This means that the attribute's mean becomes zero, and the resulting distribution has a unit standard deviation. The advantage of standardization is that it ensures that all features have the same impact on the distance metric. A prophet model has also been implemented for predicting Chicago Crime Rates. Facebook Prophet is a time series forecasting model with additive regression as its core. It works effectively with data with substantial seasonal effects and data from multiple seasons. Outliers, missing data, and rapid changes in time series are not a problem for the Prophet.

Implementation Details

One-Hot encoding:

In machine learning models, all input and output variables must be numeric. This means that categorical data must be converted to numbers before fitting and evaluating a model. One-hot encoding is a process for transforming categorical

data into a format that machine learning algorithms may use to increase prediction accuracy. When working with machine learning models, the term "one-hot encoding" comes up frequently. The sklearn handbook for one hot encoder says to "encode categorical integer features using a one-hot method." The category value represents the numeric value of the dataset entry. After that, a value of "0" denotes non-existence and a "1" indicates existence after one round of hot encoding.

Feature Scaling:

Feature scaling is an approach for standardizing the range of independent variables or data columns. It is used to handle a wide range of magnitudes of distinct columns. A machine learning algorithm will assume values which are larger are greater and smaller values are lower if feature scaling is not done, regardless of the unit of measurement. To avoid this, feature scaling is used.

There are two basic methods for scalability of features:

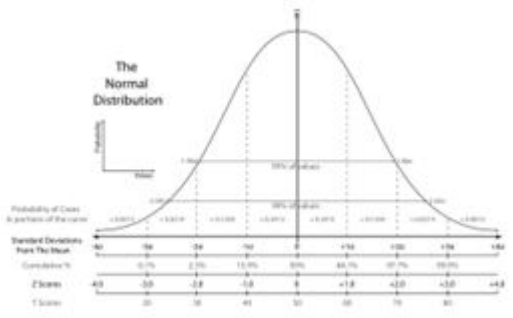
$$Z = \frac{x - \mu}{\sigma}$$

1. Normalization
2. Standardization

1. Normalization: This scaling (also known as min-max scaling) will turn the data into a range of features.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

where, x's is the normalized value and [0, 1] is an example of this. The process of translating observations into a normal distribution is known as normalization.



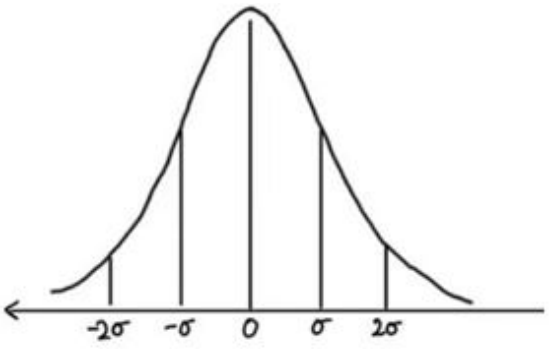
The bell curve is a statistical distribution in which roughly equal numbers of observations fall above and below the mean, the mean and median are equal, and more observations are closer to the mean.

2. Standardization:

The process of changing data into a distribution with a mean of zero and a standard deviation of one is known as standardization (also known as z-score normalization).

$$x_{new} = \frac{x_{old} - mean\ of\ x}{standard\ deviation\ of\ x}$$

where x is the original feature vector, x(mean) represents the mean of the feature vector, and s represents the standard deviation. The z- score is a statistic that is defined as



When the mean is removed from a distribution, it is shifted to the left or right by an amount equal to the mean. For example, if the mean is 100, the distribution should be shifted to the left by 100 without changing its shape. As a result, the new mean is 0. When the distribution is divided by the standard deviation, the shape of the distribution

changes. The new standard deviation for this standardized distribution is 1, which is calculated by multiplying the new mean by the new standard deviation, μ =0, where x is the original feature vector, x(mean) is the mean of the feature vector, and s is the standard deviation. Statistics provides us with the z- score, which is defined as the z-score equation.

Applications:

- In some cases, feature scaling can improve the algorithm's convergence time in stochastic gradient descent.
- It can help support vector machines discover support vectors faster.

Prophet:

The Chicago Crime dataset provides a summary of reported crimes in the city of Chicago from 2005 to 2017. In this paper, the dataset has been taken by using the API. The data was collected from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. Prophet is open-source software developed by Facebook's Core Data Science team. A prophet is a time series data prediction technique that uses an additive model to match non-linear patterns to seasonality on a monthly, weekly, and daily basis, as well as holiday impacts.

Prophet is a model that has the following components at its core:

$$y(t) = g(t) + s(t) + h(t) + \epsilon\, g(t)$$

The models' trend, which describes the long- term rise or fall of the data. Prophet offers two trend models, a saturating growth model, and a piecewise linear model, depending on the type of forecasting task. s(t) simulates seasonality using Fourier series, which displays how data is affected by seasonal variables such as the season.

Eggnog searches, for example, are up. h(t) represents a term that cannot be decreased and replicates the impact of holidays or significant events on company time series over the winter holidays (e.g. new product launch, Black Friday, Superbowl, etc.).

Prophet works best with time series with significant seasonal impacts and historical data from multiple seasons.

Design Objectives:

1. Data cleansing:

Delete or correct records with incorrect or inaccurate values, as well as records with a large number of missing columns, from raw data.

2. Instances selection and partitioning:

Data points from the input dataset are used to create training, evaluation (validation), and test sets. This includes techniques like repeating random sampling, minority class oversampling, and stratified partitioning**.**

Feature tuning:

Scaling and normalizing numeric data, imputing missing values, clipping outliers, and changing values with skewed distributions are all examples of ways to increase the quality of a feature for ML.

Representation transformation:

Using bucketization to convert a numeric feature to a categorical feature and categorical features to numeric representation (through one-hot encoding, learning with counts, sparse feature embeddings, and so on). Some models can only handle numeric data, while others can handle both. Even when models deal with both categories, having different representations (numerical and categorical) of the same data can help them perform better**.**

Feature extraction:

It can reduce the number of features by producing lower-dimension, more powerful data representations using techniques like PCA and embedding extraction. Some of the methods used include PCA, embedding extraction, and hashing.

Optional features:
Choosing a subset of the input features for training the model while discarding irrelevant or redundant ones using filter or wrapper methods. If the features misses any large number of values, just dropping them is another option.

Constructive features:
To develop extra features, traditional approaches such as polynomial expansion (using univariate mathematical functions) or feature crossover are used (to capture feature interactions). Business logic from the ML use case's domain can also be used to produce features.

Conclusion:
During data preprocessing, our designed system does data cleansing, transformation, and reduction. Our application accepts unprocessed datasets, which are then cleaned up. The cleansed data is presented to the users when all of the preparation is completed. This method saves time because manual cleaning is not required. Following purification, the user has the option of choosing or selecting a machine learning model that will provide correct plots. This is useful for users who need to clean enormous datasets and visualize the analysis of pre-processed data. Machine learning algorithm accuracy and comparison will be feasible in the future, all through a user-friendly interface. The software effectively predicted and successfully forecasted the trend of crimes committed in Chicago for the next two years, according to an investigation into how fbprophets works.

The number of crimes committed in Chicago has significantly decreased and is expected to continue to decrease, which is a good sign because it indicates that the city is becoming safer over time. The fbprophet model validates our first conclusion that crime has decreased over time, proving that it is right. While Chicago's crime rates are decreasing, they can be further reduced by implementing safety measures and ensuring that offences are properly reported.

The assertion that Fbprophet is a very convenient approach to make predictions on time series data is validated in this paper. The technique can be used to estimate crime rates in any other city in the globe because the dataset is available on the internet. With this model, anyone may discover which crimes are committed the most frequently, where they are committed the most frequently, and which crimes require immediate attention in order to lower future crime rates, assuming the dataset is publicly accessible via the internet.

References:
[1] Cristian Felix, Anshul VikramPandey, and Enrico Bertini, "textile: AnInteractive Visualization Tool for Seamless Exploratory Analysis of Structured Data and Unstructured Text", IEEE-2018.

[2] Data, Huawen Liu, Xuelong Li, Jiuyong Li, and Hicham Zhang,"Efficient Outlier Detection for High-Dimensional", IEEE- 2019.

[3] M. Bostock, V. Ogievetsky, and J. Heer, "Data Driven documents," IEEE- 2011.

[4] F. Beck, S. Koch, and D. Weiskopf, "Visual Analysis and Dissemination of Scientific Literature Collections with SurVis", IEEE-2016

[5] Parke Godfrey, JarekGryz and Piotr Lasek,"Interactive visualization of large datasets", IEEE-2016

[6] Dileep Kumar Ashley and Raju Hadler, "Data Cleaning: An Abstraction-based approach",IEEE-2015

[7] Mehmet Adil Yalçın;Niklas Almqvist; Benjamin B. Bederson, "Keshif: Rapid and Expressive Tabular Data Exploration for Novices",IEEE-2018

[8] List of United States cities by crime rate-Wikipedia

[9] https://research.fb.com/blog/2017/2/ prophetforecasting-at-scale/

[10] Data Preprocessing in Data Mining Geeks

for Geeks QuickStart | Prophet (facebook.github.io)

[11] https://medium.com/analytics-vidhy a/time-seriesanalysis-a-quick-tour-of-fb prophet-cbbfbffdf9d8

[12] Chicago Crime Rate Forecasting using FbProphet | by Eashan Kaushik https://medium.com/analytics-vidhya/chicago-crime-rate-forecasting-using-fbprophet-17757f45e9bb