CS-697AK - Exam -

WSU ID - N659A624

1) Descriptive Analysis:-

As a analyst or owner of a business, we are eagerly looking for an answer of question - What is happening in my business? This is where descriptive analytic comes into picture. This is most commonly used analytics that analyses the data coming in the real-world. We use this data in effective visualisation tools like dashboards, reports etc, which allow us to learn from Past behaviours

For example, in banking sector we need to identify the credit card fraud. Based on the transaction happening in real-world we can identify the wether the transaction is fraud or valid. and generate graph between valid and fraud transactions.

Diagnostic Analysis:-

The next step in complexity in data analytics is descriptive analytics. Here we find - Why is it happening? We need to drill down to root-cause. Ability to isolate all the confounding information. Generally in business BI dashboards helps you to drill down by doing quick comparision to find reason of factors effecting business.

Considering Credit card fraud here we identify the cause of fraud ~~may~~ Those may be security vunerable etc and list them out. It may helps to find the amount (what amount) the fraud is happening

Predictive Analysis:

Predictive Analysis is based on what you get fom descriptive and diagnostic analytics and used to find answer to question of What is likely to happen in future based on previous trends and Patterns. In general it's all forecasting. Predective analysis uses various statistical and machine learning algorithm to provide recommendations and provide answer to what likely to happen in future.

For example, for credit card fraud, we use variety of parameter like how many times (frequently) person uses credit card, in what websites of type of agents there is possibility of fraud Based on time at which transaction happend. location where it happend and location where actual person is located. Based on these we can predict the possibility of fraud.

Prescriptive Analysis:

When you get finding from descriptive, diagnosis and predictive analytics like what happened, the root cause behind that and what might happen in future. Prescriptive model utilizes those answer to help you determine the best course of action to choose to bypass & eliminate future issues.

For example, based on factor like user location, time at which transaction happened. We can contact user for verification or used OTP for more securities.

2) Feature Extraction aims to reduce the number of features in a dataset by creating new features from the existing ones. These new reduced set of features should then be able to summarize most of information contained in original set of features.

Another commonly used technique to reduce the number of features in dataset is called Feature Selection.

The difference between feature selection and feature Extraction is that future selection aims instead to rank the importance of existing features in the dataset and discard less important ones.

Feature Extraction leads to advantages like

- Accuracy Improvements
- Overfitting risk reduction
- Speed up training
- Improved Data Visualization
- Increase in explainability of our model.

For example, in credit card fraud, from the amount of data available we extract the features we required and list them.

But in feature selection we select the attributes that has high mode, mean, Variance and standard deviation so that using (selecting such features) help in increasing the accuracy of the model, speed up the training

Feature Extraction

Principle Component Analysis (PCA), is one of the most used linear dimensionality reduction technique. When using PCA we take as input our original data and try to find a combination of the input features which best summarize the original data distribution so that to reduce its original dimensions. PCA is able to do this by maximizing

Variance and minimizing the reconstruction error by looking at Pair-wised detection

PCA is an unsupervised learning algorithm, therefore it doesn't about datalabels but only about variance. This can lead in some cases to misclassification of data.

Feature Selection:

Feature Selection is crucial to any model construction in datascience. Focusing on the most important, relevant features will help any datascientist design a better model and accelerate outcomes.

Common Methods for future selection are

- Wrapper Methods
- Filter Methods.

Filter Method:-

- It generally looks at features independently, evaluating the relevance of each particular feature. It would score the features independently of how they perform on model of interest.

Wrapper Method;

It evaluates the features in relation to their performance on the model. The set of features are used to construct the model and the performance of set is scored.

3) The most widely used predictive models are:-

1) Decision Trees:-

Decision trees are simple, but powerful form of multiple variable analysis They are produced by algorithms that identify various ways of splitting data into branch-like segments. It partition data into subsets based on categories of input variables, help you to understand Someone's path of decision.

2) Regression (Linear and Logistic)

It is one of most popular methods in statistics. It estimates relationships among variables. finding a pattern in large & diverse sets

3) Neural Networks:-

Patterned after the operation of neuronsin the human brain, neural networks are a variety of deep learning technologies.

Other classifiers are

- Time Series Algorithms
- Clustering Algorithm
- Outlier Detection Algorithm
- Ensemble Model
- Factor Analysis
- Naive Bayes
- Support Vector Machines

In banking and financial Service industry, predictive analytics and machine learning are used in conjunction to detect and reduce fraud.

In credit card fraud detection, we use machine learning to detect anomolous activities called outliers. Once dataset is formatted and processed. The data is processed by set of algorithms from sklearn. This data is fit into a model and following outlier detection modules are applied on it:

- Local Outlier Factor
- Isolation Forest Algorithm.

These are part of sklearn. These include ensemble based methods and functions for classification, regression and outlier detection.

ROC Curve:

An ROC curve (Receiver Operating Characteristic Curve) is a graph showing performance of classification model at all classification thresholds. This curve plots two parameters.

- True Positive Rate (TPR)
- False Positive Rate (FPR)

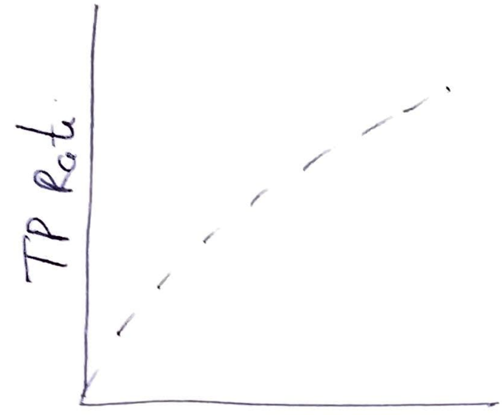
TRR is a synonym for recall and its threshold as follows.

$$TPR = \frac{TP}{TP + FN}$$

FPR is defined as

$$FPR = \frac{FP}{FP+TN}$$

An ROC curve plots TPR vs FPR at different classification thresholds. Lowering classification threshold classifies more items as positive. Thus increase in both False positives & True positives.

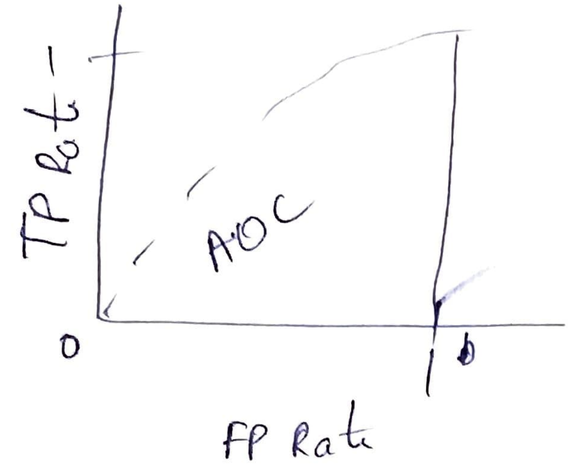


It is a typical ROC curve.

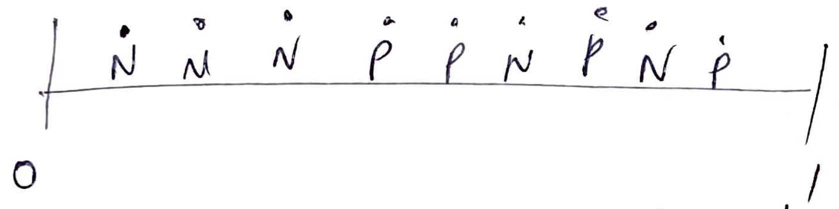
To compute points in ROC curve, we could evaluate a logistic regression model many times with different classification threshold but this would be efficient. But sorting-based algorithm that can provide information for us called AUC.

AUC (Area Under the ROC Curve)

– It measure the entire 2-D area underneath the entire ROC curve



AOC provides an aggregate measure of performance across all classification thresholds. One way of interpreting AUC is as the probability that model ranks a random positive example more highly than a random negative Example.



P – Actual positive

N – Actual Negative

AUC is desirable for following reasons.

- Scale - Invariant. – It measures how well predictions are ranked, rather than their absolute values.
- It is Classification - threshold - invariant It measures quality of model's prediction, irrespective of what classification threshold is chosen.

For credit card fraud detection, I use (for unlabeled data)

- Local Outlier factor
- Isolation forest Algorithm.

Local Outlier Factor

- It is an Unsupervised Outlier Detection algorithm. Local Outlier factor' refers to anamoly score of each sample. It measure the local deviation of sample data with respect to its neighbours. More, precisely, locality is given by k-Nearest neighbours whose distance is used to estimate the local data.

The Pseudo code for algorithm is

```
import numpy as np
import matplotlib as plt
from sklearn.ensemble import Isolation forest

 rng = np.random.Randomstate (42)
 # General Train Data

 x = 0.3 * rng.randn (100, 2)
 X_train = np.r [X + 2, X - 2]
 # Generate Some abnormal novel observation
 X_outlier = rng.uniform (low -4, high-4, size (20, 2))
```

```
# fit the model
clf = Isolation forest ( behaviour = 'new'; max-sample = 100,
                         random state = rng, contamination = 'auto')

clf. fit ( x-train)
y-pred_train = clf. predict ( x_train)
y-pred-test = clf. predict ( x_test)
y-pred-outliers = clf. predict (x-outliers)
# plot the lines, samples, and nearest vectors to the plane
 xx, yy = np. meshgrid (np. linspace(-5, 5, 50), np. linspace(-5, 5, 50))
  Z = clf. decision function (np. c_ [xx. ravel(), yy. ravel()])
  Z = Z. reshape (xx. shape)
```

On plotting the results Local Outlier algorithm we get graph. By comparing local values of a sample to that of its neighbours, one can identify samples that are substantially lower than their neighbours. These values are anomalous are they are considered as outliers.

5) In order to develop a model with high accuracy and prediction we need to know about under-fitting and over fitting.

Over Fitting:

- Over fitting refers to a model that models the training data too well.
- Over fitting happen when a model learns the detail and data noise in the training data to the extent that it negatively impact the performance of model on new data. This means that the noise of random fluctations in the training data and learned as concept by model. The problem is that these concepts do not apply to new data and negatively impact the model ability to generalize.
- Overfitting is more likely with nonparametric and nonlinear models that have more flexibility when learning a target function. As such, many non-paramatric machine learning algorithms also include parameters of techniques

Things to Overcome Overfitting

- Cross validation
- Reduce the train data
- Remove Features

Under fitting:

- Under fitting refers to a model that can ~~weigh~~ neither model training data not generalize to new data.
- An underfit machine learning model is not a suitable model and will be obvious as it have poor performance on training data.
- Under fitting is often not discussed as it is easy to detect given a good performance metric. The remedy is to move on and try alternate machine learning algorithm. Nevertheless, it does provide a good contract to problem of overfitting.

Common ways to prevent underfitting are

- Get more training data
- Add dropout
- Reduce capacity of ~~data~~ network.

→ Ideally, you want to select a model at a spot between underfitting and overfitting.

This is the goal, but it is difficult to do in practice.

Overtime, as algorithm learns the error of the model on training data goes down and so does the error on the test dataset.

If we train for too long, the performance on training dataset may continue to decrease because model is

Overfitty and learny the irrelevant detail and noise in the training set.

The sweetest spot is the point just before the error on ~~train~~ test dataset starts to increase when model has good skill on both training dataset and unseen test dataset.

The important technique you can use when evaluatny machine learning algorithm are to limit overfitty are

1) Use a resampling technique to estimate model accuracy
2) Hold back a validation dataset.

Using a cross validation is a golden standard in machine learning for estimatny model accuracy on unseen data.

If you have data, using a validation dataset is also an excellent practice.

In Credit card fraud detection we use Cross validation, limit the trainy data and followed these thing to gain the accuracy.