Program – 11

AIM: To Perform Receiver Operator Curve (ROC) analysis, Forecast analysis and Survival Analysis on the sample dataset

Introduction and Theory

Receiver Operator Curve Analysis

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection in machine learning. The false-positive rate is also known as the fall-out or probability of false alarm and can be calculated as (1 − specificity). It can also be thought of as a plot of the power as a function of the Type I Error of the decision rule (when the performance is calculated from just a sample of the population, it can be thought of as estimators of these quantities). The ROC curve is thus the sensitivity as a function of fall-out. In general, if the probability distributions for both detection and false alarm are known, the ROC curve can be generated by plotting the cumulative distribution function (area under the probability distribution from negative infinity to the discrimination threshold) of the detection probability in the y-axis versus the cumulative distribution function of the false-alarm probability on the x-axis.

ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones independently from (and prior to specifying) the cost context or the class distribution. ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making.

Survival Analysis

Survival analysis is a branch of statistics for analyzing the expected duration of time until one or more events happen, such as death in biological organisms and failure in mechanical systems. This topic is called reliability theory or reliability analysis in engineering, duration analysis or duration modelling in economics, and event history analysis in sociology. Survival analysis attempts to answer questions such as: what is the proportion of a population which will survive past a certain time? Of those that survive, at what rate will they die or fail? Can multiple causes of death or failure be taken into account? How do particular circumstances or characteristics increase or decrease the probability of survival?

To answer such questions, it is necessary to define "lifetime". In the case of biological survival, death is unambiguous, but for mechanical reliability, failure may not be well-defined, for there may well be mechanical systems in which failure is partial, a matter of degree, or not otherwise localized in time. Even in biological problems, some events (for example, heart attack or other organ failure) may have the same ambiguity. The theory outlined below assumes well-defined events at specific times; other cases may be better treated by models which explicitly account for ambiguous events.

More generally, survival analysis involves the modelling of time to event data; in this context, death or failure is considered an "event" in the survival analysis literature – traditionally only a single event occurs for each subject, after which the organism or mechanism is dead or broken. Recurring event or repeated event models relax that assumption. The study of recurring events is relevant in systems reliability, and in many areas of social sciences and medical research.

Program – 11

FORECAST ANALYSIS

Forecasting is the process of making predictions of the future based on past and present data and most commonly by analysis of trends. A commonplace example might be estimation of some variable of interest at some specified future date. Prediction is a similar, but more general term. Both might refer to formal statistical methods employing time series, cross-sectional or longitudinal data, or alternatively to less formal judgmental methods. Usage can differ between areas of application: for example, in hydrology the terms "forecast" and "forecasting" are sometimes reserved for estimates of values at certain specific future times, while the term "prediction" is used for more general estimates, such as the number of times floods will occur over a long period.

Risk and uncertainty are central to forecasting and prediction; it is generally considered good practice to indicate the degree of uncertainty attaching to forecasts. In any case, the data must be up to date in order for the forecast to be as accurate as possible. In some cases the data used to predict the variable of interest is itself forecasted.

Code

ROC

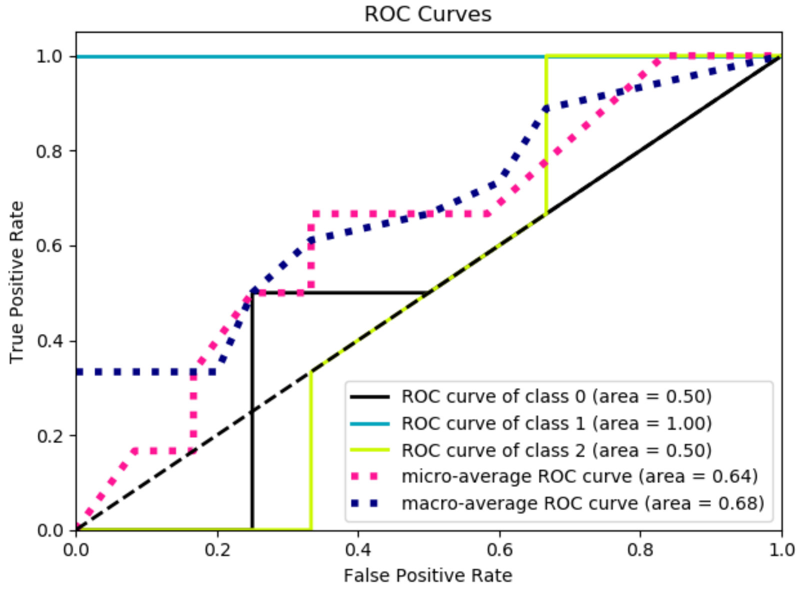
```
import scikitplot as skplt
import matplotlib.pyplot as plt
y_true = [2, 0, 2, 2, 0, 1] # ground truth labels
y_probas = [[0.1,0.4,0.5], [0.3,0.2,0.5], [0.9,0.0,0.1],
[0.2,0.2,0.6], [0.1,0.1,0.8], [0.1,0.9,0.0]] # predicted probabilities
generated by sklearn classifier
skplt.metrics.plot_roc_curve(y_true, y_probas)
plt.show()
```

Survival Analysis

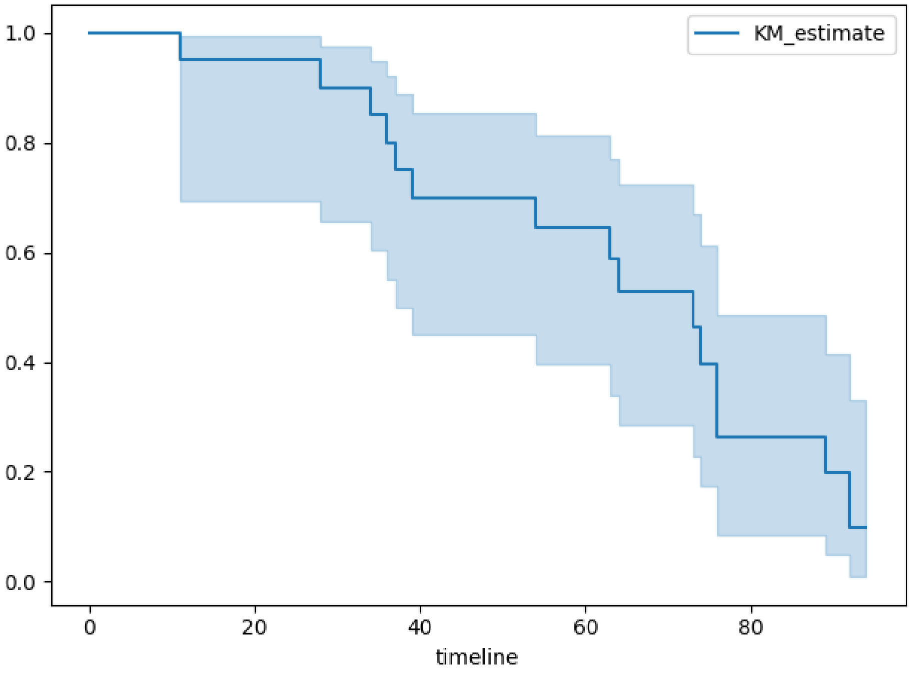
```
from lifelines import KaplanMeierFitter
import matplotlib.pyplot as plt
durations = [11, 74, 71, 76, 28, 92, 89, 48, 90, 39, 63, 36, 54, 64,
34, 73, 94, 37, 56, 76]
event_observed = [True, True, False, True, True, True, True, False,
False, True, True,
                  True, True, True, True, True, False, True, False,
True]

kmf = KaplanMeierFitter()
kmf.fit(durations, event_observed)
kmf.plot()
plt.show()
```

Program – 11

Results & Outputs



ROC



Survivial

Findings and Learnings:

1. What is ROC Curve and how ROC analysis is useful.
2. What is survival analysis and where it is used.
3. Various ways to perform forecast analysis and its importance.