Week 12
Artificial Intelligence (AI) for Investments

Lesson 2: Classification Algorithms: Application

Introduction

- Application of classification algorithm in the prediction of security prices
- Revisiting the ABC case study
- Logit/Probit modeling
- Training the model and testing the model
- Model performance evaluation
- Summary and concluding remarks

Case Study: ABC Stock Price Forecasting

Case Study: Stock Price Prediction

- Stock price prediction or stock return prediction is an attempt to determine the future value of a company based on an analysis of factors, which impact its price movement
- There are a number of factors that help in predicting stock prices
- These can be macroeconomic factors like the state of the country's economy, growth rate inflation, etc.
- There are also other factors that are more specific to a stock like profit margin, debt to equity issues, sales of a company, etc.

Case Study: Stock Price Prediction

So, we are given the data for stock market price for ABC company, along with Nifty and Sensex (market indices). We are also given the data of dividend announcement and a sentiment index.

Date	Price	ABC	Sensex	Dividend Announced	Sentiment	Nifty
03-01-2007	718.15	0.079925	0.073772	0	0.048936	0.095816
04-01-2007	712.9	–0.00731	0.021562	0	–0.05504	0.009706
05-01-2007	730	0.023987	–0.02441	0	0.019135	–0.03221
06-01-2007	788.35	0.079932	0.012046	0	0.080355	0.011205
07-01-2007	851.4	0.079977	–0.0013	0	0.094038	–0.0004
10-01-2007	919.5	0.079986	0.019191	1	0.015229	0.030168
11-01-2007	880	–0.04296	–0.04025	0	–0.07217	–0.04966
12-01-2007	893.75	0.015625	0.036799	0	0.01396	0.020999
13-01-2007	875	–0.02098	–0.00845	0	0.057518	–0.01164
14-01-2007	891	0.018286	0.004858	1	0.008828	0.020714
17-01-2007	819.75	–0.07997	–0.01228	0	–0.12395	–0.00962
……	……	……	……	……	……	……
……	……	……	……	……	……	……

Case Study: Stock Price Prediction

- Consider a portfolio manager who has built a model for a particular stock
- The manager wants to predict whether in the next period the ABC stock price returns for this stock will go up or down
- The data starts from 2007 and goes till 2019, so we have approximately 13 years of data
- We have daily returns of ABC or a change in the price of ABC in column B. Next, we have a daily return on Sensex in column C and a daily return on Nifty in column D.

Case Study: Stock Price Prediction

- Sensex and Nifty are the two main stock indices used in India
- They are benchmark Indian stock market indices that represent the weighted average of the largest Indian companies
- So, Sensex represents average of 30 largest and most actively traded Indian companies
- Similarly, Nifty represents a weighted average of 50 largest Indian companies

Summary

The following tasks need to be performed

- Create a dummy variable that is 1 when stock prices go up and create a dummy variable that is “0” when stock prices go down
- Segregate the data into test and train datasets
- Train and build the model using simple logit/probit classification algorithms using market index as the independent variable, and up/down dummy as the dependent variable

Summary

The following tasks need to be performed

- Evaluate the in-sample performance and out-of-sample performance of the model
- Compute the marginal effects of the independent variable
- Visualize the performance of these models using the ROCR curve
- Examine the classification accuracy of the model and compare it with a similar linear probability model
-

Data Input and Exploration

Data Input and Exploration

- In this video, we will start with the implementation of the classification algorithms using ABC Case study Data
- First, we will set the working directory, then we will read the data
- Lastly, we will create the binary response variable: ‘1’ for positive returns and ‘0’ for negative returns

Summary

- We started our analysis with setting the working directory
- Next, we loaded the relevant package libraries
- Then we read the data from the working directory
- Lastly we created a new ‘updown’ binary response variable, which is ‘1’ when returns are positive and ‘0’ when returns are negative

Creation of Test and Train Datasets

Creation of Test and Train Datasets

- In this video, we will create the test and train sample datasets
- Then we will examine the distribution of our binary response variable in 1's and 0's

Summary

- First, we filtered the observations after 2006 and cleaned our data
- Next, we randomly selected 80% observations as training dataset and remaining 20% as test dataset
- Lastly, we tested the proportion of 1's and 0's in the parent dataset, test dataset, and train dataset
- The distribution of 1's and 0's is fairly similar for all the three datasets

Training the Linear Probability Model (LPM) Algorithm

Training the LPM Algorithm

- In this video, we will train an LPM algorithm with the training dataset
- Next, we will compute the classification/confusion matrix
- Final, using the classification/confusion matrix, we will compute various performance measures, i.e., accuracy, specificity, and sensitivity

Summary

- We trained an LPM algorithm using the training dataset
- Using the fitted results, we converted them into 1's and 0's using thresholding values of 0.4, 0.6, and 0.8
- Lastly, using the classification/confusion matrix, we computed three performance parameters, namely, accuracy, specificity, and sensitivity

Training the Logit/Probit Algorithms

Training the Logit/Probit Algorithms

- In this video, we will train the Logit/Probit classification algorithms using the training dataset
- Next, we will compute the in-sample performance evaluation measures
- We will also compute the marginal effects of the independent variable on the dependent variable
- Lastly, we will evaluate and compare the performance of these algorithms on parameters, namely, accuracy, specificity, and sensitivity

Summary

- We trained our classification algorithms using the training dataset
- Next, we computed the Pseudo R-square measure and also computed the marginal effects
- Lastly, we evaluated the performance of these algorithms on three parameters of sensitivity, specificity, and accuracy, using classification matrix at threshold values of 0.4, 0.6, and 0.8
- The performance of all the algorithms appear to be close to each other; this is ascribed to the fairly symmetric distribution of 1's and 0's in the training dataset

Visualizing the Performance

Visualizing the Performance

- In this video, we will compare the performance of the three trained classification algorithms (linear, logit, and probit objects) using correlation measure and through visualization

Summary

- We computed the correlation across the fitted values for the three classification algorithms (linear, logit, and probit)
- The correlations appear to be very high
- Next, we visualized the performance of the algorithms on parameters of accuracy, sensitivity, and specificity for the three threshold values of 0.4, 0.6, and 0.8
- While the performance of these algorithms appear to be close, logit model appears to offer the best fit, followed by the probit, and then the linear model

Receiver Operating Characteristic (ROC) Curve

ROC Curve

- In this video, we will compare the performance of the three trained classification algorithms (linear, logit, and probit objects) with the help of ROC curve

Summary

- We plotted the ROC curve and examined the performance of the three trained classification algorithms
- Area under the curve (AUC) appears to be identical for all the three algorithms; this is ascribed to the extremely high correlation in the fitted objects of these models

Defining the Objective Performance Function

Defining the Objective Performance Function

- In this video, we will develop a simple machine learning system that will help the computer learn how to select the best classification algorithm across a class of algorithms
- We will create a suitable user defined performance function to analyze the performance of these algorithms

Summary

- We created an optimization function, which included the arguments, namely, fitted values, actual values, and simulated threshold values
- These values are employed to compute accuracy, sensitivity, and specificity parameters through classification matrix
- The final performance object is a simple average of these three parameters (i.e., accuracy, sensitivity, and specificity)
-

Creating Performance Objects

Creating Performance Objects

- In the previous video, we defined our performance objective function; in this video we will simulate 1000 threshold values and calculate the performance object values for all the three classification algorithms using these threshold values

Summary

- We created three performance objects for the three classification algorithms, namely logit, probit, and linear
- We simulated 1000 performance object values using our performance objective function for all the three algorithms (linear, logit, and probit)

In-sample Performance Evaluation

In-sample Performance Evaluation

- In the previous video, we computed 1000 performance object values for the three classification algorithms
- In this video we will compare the performance of these three classification algorithms through visualization

Summary

- We plotted 1000 performance object values for our three classification algorithms, namely, linear, logit, and probit
- We found that for most of the threshold values, the logit model algorithm works best, closely followed by probit model algorithm, and lastly the linear model algorithm
- Lastly, we extracted the best fit model and the corresponding threshold value

Out-of-Sample Prediction

Out-of-Sample Prediction

- In this video, we will start with out-of-sample prediction
- We will use the trained algorithms for our linear, logit, and probit models and predict using test data set
- Lastly, we will compute the correlations across the predicted values between the three algorithms

Summary

- We performed the prediction on the test data using our trained algorithms for linear, logit, and probit models
- We found that the correlation across the predicted values are very high; in fact the correlation between logit and probit predicted values are 99%, and the correlation with linear model predicted values are more than 90%
- This is ascribed to the fact that correlations across predicted objects are very high, and the distribution of 1's and 0's is highly symmetric in our test and training datasets

Out-of-Sample Prediction: ROC Curve

Out-of-Sample Prediction: ROC Curve

- In the previous video, we performed prediction with trained algorithms, using the test datasets
- In this, video, we will visualize and compare the performance of the three trained algorithms, using ROC curve and also compute area under the ROC curve

Summary

- We plotted ROC curves for all the three classification algorithms for linear, logit, and probit models
- The performances as per the ROC curve are quite similar with identical area under the curve (ROC)
- This is ascribed to the high correlation across fitted objects and symmetric nature of 1's and 0's in our test and training datasets
- In the next video, we will simulate 1000 threshold values and compute the performance object values

Out-of-Sample Prediction: Performance object

Out-of-Sample Prediction: Performance object

- We have already set-up a performance object, which is the average of three parameters: accuracy, sensitivity, and specificity
- Using our predicted values for all the three algorithms, we will compute the performance object values for the 1000 simulated threshold values

Summary

- In this video, we computed the values of our performance object using 1000 simulated threshold values for all the three algorithms, i.e., linear, logit, and probit
- In the next video, using these values of the performance object, we will visualize and compare the out-of-sample performance of the three algorithms

Out-of-Sample Prediction: Performance Evaluation and Visualization

Out-of-Sample Prediction: Performance Evaluation and Visualization

- In the previous video, we have simulated 1000 performance object values using our trained algorithms with the test data
- In this video, using these performance object values, we will visualize and compare the performance of the three trained algorithms

Summary

- To summarize, we plotted our simulated performance object values
- For most of the threshold region, the logit model offers the best prediction, closely followed by the probit and linear models
- We also extracted the details corresponding to the best performance object value, including its threshold level

Summary and Concluding Remarks

Summary and Concluding Remarks

- ABC stock price up/down movements are modelled using logit/probit classification algorithms
- The model is trained using the training dataset and is examined on various measures of model performance evaluation
- Fitted modelled is examined visually as well

Summary and Concluding Remarks

- The model is tested using test dataset and various measures of out of sample fit are examined
- Marginal effects of these independent variables are computed
- The performance of this model is compared with a similar linear probability model

Thanks!