**Trimester: FIFTH TRIMESTER**

**Name(s): ANAND MOHAN  
SRN: R19MBA53**

|  |  |
| --- | --- |
| Project Name: | Regime Prediction using Machine Learning in Stock Market |
| Background Information | Majority of share market investment decisions are highly influenced by herd mentality. To avoid herd mentality, we should not blindly invest in Assets based on what others are doing.  When making investments in Asset classes, we need to study the related factors more carefully. In some industries, data is not easily available and we need to put in extra efforts to find and evaluate data and make wiser decisions accordingly.  Similarly, balance sheets can be manipulated by the management. Therefore we should not invest in Assets based on their face value; rather we should do ample research and then invest.  Many investors lose money in the share market due to fear, greed and inability to control their emotions.  In the bull market, investors try to speculate on assets and create heavy positions in greed to earn more and end up in losses. In the bear market, investors panic and sell their Assets out of fear and incur huge losses. By having a practical, non-emotional approach and Analyzing data by applying descriptive, predictive and prescriptive Analytics, we can create wealth in a healthy manner.  Even though we may be committed to sticking with our investing strategy, we will still need to check in periodically and make adjustments. We must do an in-depth review of our Asset portfolios on a quarterly basis. While we may not need to check in quarterly if we are passively investing in Asset classes, most advisors recommend at least an annual check in.  Since the share market doesn't behave consistently at all times, identifying the right time to invest in them can be most beneficial. Buying the identified Assets at the lowest price level will increase the potential profits investors can earn. Moreover, selling the Assets when it is trading at the highest price is profitable. |
| Statement of the Problem | The key determinant of whether the market is bull or bear is not just the market's knee-jerk reaction to a particular event, but how it is performing over the long term. Small movements only represent a short-term trend or a market correction. Whether or not there is going to be a bull market or a bear market can only be determined over a longer time period.  However, not all long movements in the market can be characterized as bull or bear. Sometimes a market may go through a period of stagnation as it tries to find direction. In this case, a series of upward and downward movements would actually cancel-out gains and losses resulting in a flat market trend.  It is a known fact that investor psychology and sentiment affect whether the market will rise or fall. Because the businesses whose Asset classes are invested on the exchanges are participants in the greater economy, it is already known that the stock market and the economy are strongly linked. Based on these known factors, Market analysts come up with their own fundamental Analysis with some key Market ratios and Technical indicators. This has been already done earlier which roughly give an estimate of the Bullish and Bearish Regimes.  However, a requirement comes for us to be able to estimate and benchmark our accuracies in able to predict Bullish and Bearish regimes by using the Asset class returns as Independent variables and able to predict Regimes period (Bullish or Bearish) to identify Recession periods spot on.  My major milestone targets would be trying simpler and easy to understand solutions which can get most effortlessly applied by any non-professional investor and which still helps them to optimize their portfolios decisions based on Bullish and Bearish Regimes. |
| Proposed Solution | We will employ different Modelling algorithms on the transformed data which mainly comprises of Returns values for different Asset classes. We will be using Regime as Target or Dependent Variable to predict bullish or bearish Regimes. Initially we will apply Logistic Regression Classifier by tuning hyper parameters such as Logistic Regression with no penalty, L1 penalty, and L2 penalty. We will also be tuning hyper parameter Inverse regularization parameter C differently in Logistic Regression. Regularization is applying a penalty to increasing the magnitude of parameter values in order to reduce over fitting. Then we will apply Nonlinear Classifier Models namely Decision Tree Classifier Modelling, Random Forest Classifier Modelling and XGBoost Classifier Modeling.These Modelling algorithms will be applied several times by tuning their hyper parameters differently. We will use 4 different Evaluation Metrics namely classification accuracy (ACC), quadratic probability score (QPS), Matthew's Correlation Coefficient (MCC) and Area under the ROC curve to determine the quality of predicted outputs.   We will be applying Cross-Validation Technique and Rolling-Window Technique while Building prediction Models. |
| Detailed Scope of Work: | Assumptions of Time series Model are that Data should be stationary. Non stationary data may show the trend. But we should convert non stationary data to stationary data to do better forecasting. We will convert the features into stationary form by applying the necessary transformations namely Differencing and Log transformation. Then we will Add 1, 3, 6,9,12 months lags of the features.  Then we will be employing different Modelling algorithms on the transformed data. We will be using Regime as Target or Dependent Variable to predict bullish or bearish Regimes.  FRED is an online database which provides me the actual returns data for the Asset classes. It also provided Regime\_index1.The index is a benchmark market index that represents the weighted average of all the Asset classes’ returns.  Based on Index trend, Bullish or Bearish Regime is determined.  Based on Bullish or Bearish trends, Regime\_index1 marks the Bullish Regime as 0 and Bearish period as 1.Similarly Regime\_index2, Regime\_index3, Regime\_index4, Regime\_index5 is obtained.  We will be using Logistic Regression modelling several times by tuning hyper parameters differently by varying hyper parameter penalty as None, l1 and l2 and also modifying Inverse regularization parameter C differently as 0.0001, 0.01, 0.1, 1, 10 and 100.For the Logistic Regression Modelling we will also be defining the hyper parameters solver as 'saga'and max\_iter as 100.  We will be using Decision Tree Classifier Modelling several times by tuning hyper parameters differently such as max\_depth as 3, 5, 8, 10,splitter as best and random,min\_samples\_split as 2, 3 and 5.  We will then also be using Random Forest Classifier Modelling several times by tuning its hyper parameters differently such as random\_state as 42, max\_depth as 3, 5, 8 and 10,n\_estimators as 100,200 and 400.  We will then also be using XGBoost Classifier Modelling several times by tuning its hyper parameters differently such as booster as gbtree,max\_depth as 3, 5, 8 and 10,n\_estimators as 100,200 and 400,random\_state as 42,objective as binary: logistic.  Then we will be determining the best solution possible from each of the six different types of Modelling Algorithms namely Logistic Regression modelling with penalty as Null, Logistic Regression modelling with penalty as L1, Logistic Regression modelling with penalty as L2, Decision Tree Classifier Modelling, Random Forest Classifier Modelling and XGBoost Classifier Modelling.  We will be using 4 different evaluation Metrics namely classification accuracy (ACC), quadratic probability score (QPS), Matthew's Correlation Coefficient (MCC) and Area under the ROC curve to determine the quality of predicted outputs.  The model outputs are received in two forms: probability forecasts and binary classification of regime states. We consider multiple error metrics to evaluate a classification model performance. A natural metric is classification accuracy (ACC).  Quadratic probability score (QPS) evaluates prediction performance in terms of probabilities. QPS does not evaluate classification ability. If two models have different probability outputs, they might have same ACC score but not same QPS which is mean squared error of probability forecasts. In a perfect prediction setting, QPS equals 0. The lower the value of QPS, the better is the model performance.  The Matthew's Correlation Coefficient (MCC) is also reported as an evaluation metric and deployed widely in the field of computational biology. In perfect prediction setting MCC equals 1 and the worst value is -1. Therefore the higher the MCC score the better the classification performance.  Receiver operating characteristic (ROC) curve is a common choice in imbalanced classification problems. ROC curve plots true positive rate against false positive rate at various threshold settings. Area under the ROC curve (AUC) generates summary statistic for ROC metric. Higher the AUC value, the better is the model performance.  We will be applying Cross-Validation Technique and Rolling-Window Technique while Building prediction Models.Cross-validation is a common framework for hyper parameter tuning in the ML field. In the standard k-fold framework, training data is randomly grouped into k folds. In each iteration, model is trained on k-1 folds, and the remaining fold is used for validation. After k iteration, average of the model score is computed from validation sets.  Since our dataset has a time series property, we implement the cross-validation to keep temporal dependency and avoid look ahead bias. We create  𝑘 -folds as block of time periods, and move training and validation sets in a rolling basis. After the model hyper parameter optimization, out-of-sample predictions are performed in a rolling window basis with a length of 150 months.  Finally we will be able to determine the Modelling Algorithms giving the most accurate predictions about Target variable namely Regime to predict a period as Bullish (0) or Bearish i.e. Recession period (1) and hence clearly define and pin point the Recession zones when either we can cut down on our investments in Asset classes or continue to hold on to our existing investments without making any newer investments in the Recession zones.  The objective is to survey, study, and examine various facets and provide better solutions to predict Bullish and Bearish Regimes while investing in Asset Classes. |
| Support needed from Program office | Yes. Support needed from Program office for more ideas and better implementation. |
| References | [1] Arturo Estrella and Frederic S. Mishkin , PREDICTING U.S. RECESSIONS: FINANCIAL VARIABLES AS LEADING INDICATORS  [2] Periklis Gogas ,Dionisios Chionis, Ioannis Pragkidis(March 2009), Predicting European Union recessions in the euro era- Democritus University of Thrace  [3] Serena Ng,Department of Economics, Columbia University,Viewpoint: Boosting Recessions  [4] Melody Y. Huang, Randall R. Rojas, Patrick D. Convery, Department of Economics,University of California, Los Angeles(May 30, 2018),News Sentiment as Leading Indicators for Recessions  [5] Olivier Jean Blanchard and Stanley Fischer(March 10-11, 1989),NBER Macroeconomics Annual 1989, Volume 4  [6] Ruslan Bikbov and Mikhail Chernov(November 11, 2008),Monetary Policy Regimes and The Term Structure of Interest Rates  [7] Burak Saltoglu,M. Ege Yazgan(November 2012),The Role of Regime Shifts in the Term Structure of Interest Rates: Further Evidence from an Emerging Market  [8] Jonathan H. Wright(2006-07),The Yield Curve and Predicting Recessions  [9] Trevor Hastie,Robert Tibshirani,Jerome Friedman,The Elements of Statistical Learning |