Cute2 Assignment

By Harsha, Vinod & James

INSOFE - Batch31

Agenda

- Requirement Definition
- Study & Analysis
- Approach
- Model Building
- Testing Models
- Conclusion

The Ask : Requirement Definition

Stock market prediction Engine for algorithmic trading

Volatility Classification

Predict the volatility of the financial asset given the historical daily data related (though anonymized)

In our case y2

Quotidian Prediction

Given the daily %change in the financial asset predict %gain or %loss based on the historical data

In our case y1

Study & Analysis

Givens

- ✓ Anonymized data
- ✓ Train and Test data provided

Missingness

- ✓ Anonymized –unable to apply any domain or industry knowledge
- ✓ Lots of missing data
- ✓ Timestamp given in sequential days

Train Data
Rows : 1769
Cols : 111
NAs : 874

Test Data
Rows : 30
Cols : 109
NAs : 0

- ✓ Comparatively decent data
- ✓ Values small / fractional
- ✓ Too many columns

Data Analytical Approach –

Requirement Understanding & Data Gathering
Data Preprocessing (Impute, standardize…)
Prepare Train, Validate and Test Datasets

Choose Modeling approach
Create different models (checking significance etc)

Model evaluation – Apply model on validate data set and check results
Report on performance – accuracy / RMSE etc

Deploy Model on Test – PROD in this instance
Examine the values and check for any aberrations

Classification Problem

Prediction of Y2: Volatility Binary Index

Preprocessing

- Initially we processed train data but data set was small and therefore we combined test data to impute missing values and standardize the data
- Of course we split the data set later on to get back original train and test data sets
- Missing Values : Used central imputation as knn imputation was not working due to non-availability of neighbours in many instances. Verified that missing values were none.
- Used decostand for standardizing (Z standard).
- Additionally we did a count of NAs in columns to see if we can eliminate some columns but no luck
- Checked if we can eliminate any rows wherein we had many columns having missing values – none found eligible

Classification Problem

Prediction of Y2: Volatility Binary Index

Modeling & Evaluation

- GLM – logistic regression did not output any sensible model as it chose many variables. The residuals range was also high in the +/- 8 range. The AIC was **9226.9**
  We observed that almost all attributes were flagged significant. This will not work.
  We need to reduce dimensions.

- Tried PCA but the R-squared value was very low, so discarded this approach.

- Invoked stepAIC for feature selection. stepAIC optimized at AIC :**7204.47**

- Validated stepAIC model to predict the Train values and observed accuracy @**91**%
- Checked the residual plots and tried to process outliers by omission but couldn't

- Validated stepAIC model on the "validate" data set and observed accuracy @**90**%
- Plotted ROCR curve to predict the threshold values. We observed the curve and took a judgement call to have threshold at 0.55

Classification Problem

Prediction of Y2: Volatility Binary Index

Evaluation

- Validated stepAIC model on the “validate” data set and observed accuracy @**90**%
- Plotted ROCR curve to predict the threshold values. We observed the curve and took a judgement call to have threshold at 0.55. Of course we iterated with a few threshold values and narrowed down to 0.55.

Custom confusion matrix

```
> pred<-predict(step_mod,newdata=val.cl, type = "response")
> tab<-table(pred>0.5,val.cl$y2)
> accu<-sum(diag(tab))/sum(tab)
> accu
[1] 0.9159664
> tab

         0   1
  FALSE 211   6
  TRUE   34 225
> accu
[1] 0.9159664
>
```

Confusion matrix

```
> confusionMatrix(a,val.cl$y2)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 211   6
         1  34 225

               Accuracy : 0.916
                 95% CI : (0.8873, 0.9393)
    No Information Rate : 0.5147
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8324
 Mcnemar's Test P-Value : 1.963e-05

            Sensitivity : 0.8612
            Specificity : 0.9740
         Pos Pred Value : 0.9724
         Neg Pred Value : 0.8687
             Prevalence : 0.5147
         Detection Rate : 0.4433
   Detection Prevalence : 0.4559
      Balanced Accuracy : 0.9176

       'Positive' Class : 0
```

Classification Problem

Prediction of Y2: Volatility Binary Index

Deployment

- Validated stepAIC model on the “validate” data set and observed accuracy @**90**%
- Plotted ROCR curve to predict the threshold values. We observed the curve and took a judgement call to have threshold at 0.55. Of course we iterated with a few threshold values and narrowed down to 0.55.

- Deployment: Test data set was the deployment stage for us. Applied model on test data and output the predictions to the predictions.csv. Did a visual sanity check and observed that the values were similar to the neighbours in the train/validate data sets

Linear Regression Problem

Prediction of Y1: Price percentage change

Preprocessing

- To avoid redundancy, we state that the processes followed are similar to the logistic regression.
- We tried to include the classification variable, y2 but dropped this idea due to time constraint. Would have been good, to have tried this out and check out the regression.

10809.5

Linear Regression Problem

Prediction of Y1: Price percentage change

Modeling & Evaluation

- We tried timeseries but were not able to see clearly the trend, seasaonality and cyclicality.
- We tried with simple LM but there was very low significance and R-squared value was very low.
- Tried PCA but the R-squared value was very low, so discarded this approach as well.
- R-squared : 0.137
- After different trials we Invoked stepAIC for feature selection. stepAIC optimized at AIC :**10809.5**

	mae	mse	rmse	mape
With LM alone on Train Data	0.011972004	0.000260621	0.016143746	3.405703428
Prediction after stepAIC on Train	0.013866183	0.000337196	0.018362899	**2.026710003**

MAPE reduced

Open Questions

- Are we allowed to combine Train and Test. Test should be unseen right?
- How to map outliers to the data file

Great initiative. This cute has helped us learn so much. Of course time is always a constraint.