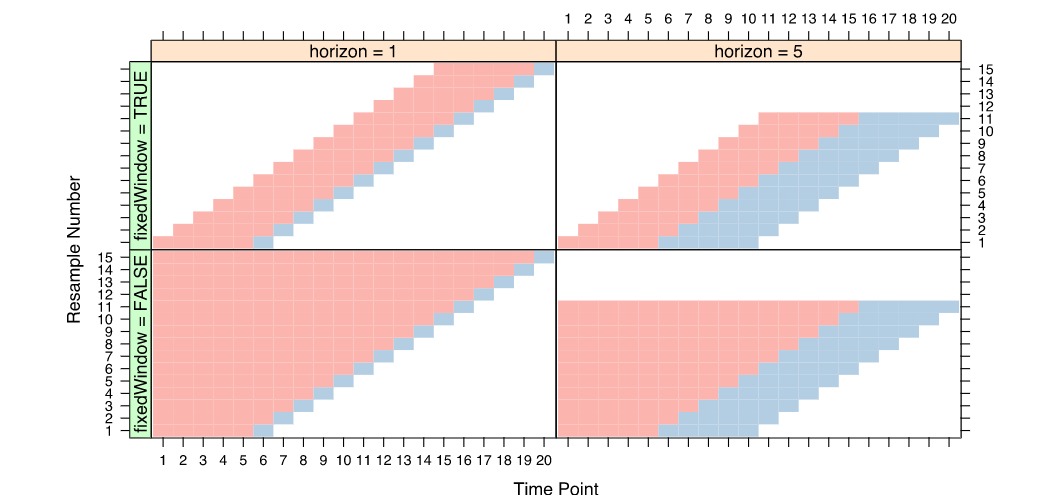
Most important thing was to have a good cross validation strategy. Task was to predict if user will buy the product in next 3 months or not. Time series CV strategy was used. Following are ways to perform time series CV.



I have used caret package for implementing this CV. I have not used any pre-processing because I was using tree based methods instead of linear and its variants methods which are effected by pre-processing. I tried some new features like difference in created date and signup date, removing factor labels which are very small in proportion.

Final model is ensemble of random forest on different cut offs for F1 score. Also as I was using parallel processing which may not suitable for reproducing the results. This can also happen because of small dataset which leads to large variance in results.

I wanted to try SVM and elastic net, xgboost and NN, but due to machine constrains I can’t. I tried up sampling, down sample and smote, which didn’t work. Also these synthetic data generation process does not work well when it comes to generalization, which are often seen when we work on heterogenous data, which can be shown by better scoring methods compare to F1, AUROC like brier score.