My experiments with Big Data

This blog is about my learnings in big data, product management and digital advertising.

Thursday, October 3, 2013

Benchmark Bond Trade Price Challenge - Kaggle

This post was long overdue. I participated in the benchmark trade bond pricing challenge and used a regression based approach to predict bond prices. Here is an outline of the approach.

1. Build on the training set and predict on the test set. The dependent variable we are trying to predict is the bond price and the independent variables are last 10 trade prices.
2. Prepare frequency charts based on callability is 0 or 1.
3. Divide the data into 12 parts – callability, price > 100 or price < 100(bond price will always converge to 100 so the curve will look different), types of trade in the bond (dealer to dealer, dealer to client, client to client – quotes driven market)
4. Some of the values were missing – missing value treatment (based on exponential weights)
5. Run regression on these sub-data sets and analyze the results
6. Some of the t-tests failed – bond ids and time to delay – p value that was kept as cut off for rejecting the coefficients is 3%

Notes :

1. R^2 is a statistic that will give some information about the goodness of fit of a model. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression line approximates the real data points. An R^2 of 1.0 indicates that the regression line perfectly fits the data.
2. R^2 measures goodness of fit. But it will not detect overfit because it will increase with any new predictor (unless it has already reached 1).
3. Assumptions of Regression
4. L Linear relationship
   I Independent observations
   N Normally distributed around line
   E Equal variance across X's
5. Multicollinearity : When two independent variables are correlated and its detection – Variance Inflation Factor
6. A p-value is the probability of an observed or more extreme result arising by chance. So, if p<.03, then that probability is quite less and hence, we can keep that independent variable

I have been away from data science now because of job change and interviewing. New Year Resolution : get back to kaggle .

Posted by Sayantan Ghosh at 5:44 PM

Popular Posts

How to score your model using scoring functions in Python
The scoring parameter can be a that takes model predictions a truth. However, if you want to scoring function th...

When to use logistic regression logistic regression
The general logistic regression does not work very well for sm set. The general logistic regres is describe...

Benchmark Bond Trade Price Cl Kaggle
This post was long overdue. I p in the benchmark trade bond challenge and used a regressic approach to predict...

Decision Trees tips

Kaggle Solutions
Dont overfit : https://www.kaggle.com/c/ov ums/t/593/results-auc Predict biological response : https://github.com/emanu...

Follow by Email

Email address...

Eyeballs

2,588

Blog Archive

▶ 2014 (46)
▼ 2013 (19)
▶ December (9)
▶ November (3)
▼ October (7)
Tim Minchin's speech at UW
Product Management and Bi : Cool featur...
What does the US governme