Bond Liquidity Prediction

Team Renegades

Methodology Followed:

Data pre-processing, Understanding the dataset, Building a suitable model and Applying to the cleaned dataset.

Data at a Glance:

There are 17261 bonds coded as 'isinyyy'. The Different characteristics of these bonds was given as the ML_bond_metadata.csv file. The data pertaining to the trade of the bonds was given in dataset.csv file. Dataset is relatively dense with near complete columns.

Data Pre-Processing:

The different type of variables were identified. Imputing data is the next step in the process. Categorical data was imputed with the mode of the column whereas numeric columns were imputed with the median values. This was done by using in-built functions in scikit-learn module of Python Programming Language.

Clustering:

The metadata file contains various categorical features as well as numeric features. To get a better understanding of the data, clustering was to be done. To gain better insights two new features were added to the data which are namely: 'AmtDiff' and 'DateDiff' which show the difference in the amounts in the market and duration of validity of the bond. The data was also scaled using StandardScaler() function in 'scikit-learn' module of Python. Traditional clustering algorithms either work for numeric or categorical data alone. But since this problem contains both categorical as well as numeric data, a novel method k-prototype clustering is used. This methods unique feature is that it takes binary distance for categorical data and Euclidian data for numeric data. The number of clusters chosen are 10. This algorithm has been coded in the 'kmodes' module of Python.

Modelling:

Observing the data reveals a peculiar nature of the data. Different data points have values at different time instances and not necessarily the same number of time points. This falls under the category of **Longitudinal Data**

Longitudinal Data:

Also referred to as panel data, track the same sample at different points in time. In contrast, repeated cross-sectional data, which also provides long-term data, gives the same survey to different samples over time. Longitudinal data have a number of advantages over repeated cross-sectional data. Longitudinal data allow for the measurement of within-sample change

over time, enable the measurement of the duration of events, and record the timing of various events. It is a growing field with continuous developments.

The sparseness of the data along time, dropping of the test subjects along time etc. give rise to difficulties in modelling this type of data. The most widely used models for the analysis of longitudinal data are mixed-effects regression models and generalized estimating equation (GEE) models.

Mixed-Effect Regression Models: These models are also called full-likelihood methods. They make full use of data of each subject. They are better than traditional ANOVA and multivariate growth curve because of the missing data. A wide range of variants of the Mixed-Effect Regression Models have been developed namely, random-effects models, variance component models, multilevel models, two-stage models etc.

Advantages of MRM models are:

i) Subjects are not assumed to be measured on the same number of time-points; thus, subjects with incomplete data across time are included in the analysis.

ii) Both time-invariant and time-varying covariates can be included in the model.

iii) Finally, whereas traditional approaches estimate average change (across time) in a population, MRMs can also estimate change for each subject. These estimates of individual change across time can be particularly useful in longitudinal studies where a proportion of subjects exhibit change across time that deviates from the average trend.

For the purpose of modelling the given data, we have chosen a Linear Mixed Effect Model. We have used 'statsmodel' module available in python programming language for the purpose. The inbuilt function MixedLM allows us to define the exogenous and endogenous variables separately. The function has a fit() method which fits the given data to the function given. The function is the dependency between the exogenous and endogenous variables. This algorithm also allows us to group data based on a certain parameter. It is especially useful to define dependency within a given group.

For the purpose of this question all the buy/sell transactions in a particular day are combined into one single buy/sell transaction. The date is taken to be the time varying parameter.

Modelling was done on an iterative basis to determine which features should be present and which shouldn't.

The predicted values are then processed in excel to give the final csv.

References:

1. Advances in Analysis of Longitudinal Data, Robert D. Gibbons, Donald Hedeker, and Stephen DuToit, US National Library of Medicine National Institutes of Health.

2. Clustering large data sets with mixed numeric and categorical values, Huang, Z, Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference, Singapore, pp. 21-34, 1997.