CLASSIFICATION MODELLING USING DECISION TREE

Bonnie Bernard

Student, PGDM-Business Analytics

Senior Manager - Transformation Lead

TATA Consultancy Services

REVA University

Bangalore, India

bonnieb.ba01@reva.edu.in

ABSTRACT

One of the operational challenges that the Investment Banks are facing today is to maintain a highly accurate and consistent data across various systems and databases. The Investment Banks are generally slow in adapting to a newer a more robust platforms and systems. Though there are several such next-gen platforms offered by the service provides in the market, due to high cost and legacy information these transitions will span over anywhere between three to five years. As a simple use case of onboarding a new customer onto the Bank's trading platform would require an average six to seven different downstream systems. And all this is done manually. Though in the recent few years, Investment Banks are opening up to the concept of Robotic Process Automation and in some cases leveraging Artificial Intelligence, the potential for big data analytics is very prominent in the coming days and years.

This paper intends to cover a simple use case for predicting transactional errors while processing for an on-boarding cum fulfilment function across various products like Equities, Fixed Income, Forex and Money Markets and Over the Counter trades.

The data set is the records of transactions performed by the team for the last 6 months with about 11 variables. The predictor variable is 'Returnflag', 0=FALSE/1=TRUE and the data is an imbalanced data, where the class of '1' is significantly lower (0.75%).

Due to its simplicity and ease of understanding by the non-analytics stakeholders i.e. the operations folks, decision tree algorithm is used as the technique for building the classification model. The benefit of identifying potential bad transactions will help the process to control a number of bad transactions to be flown into the system. This will significantly reduce remediation effort and improve first pass yield, which is one of the key performance indicators of the process.

Keywords: *Error Prediction, Decision Tree, classification model for imbalance data, over and undersampling*

CLASSIFICATION MODELLING USING DECISION TREE

1.0 INTRODUCTION

New client onboarding is emerging as the first and foremost focal point for banks and financial institutions. It is one of the most effective sales strategies which will result in better customer engagement and retention (Cognizant, white paper, 2013). The following are the key process performed while on-boarding a new client, Know Your Customer, Anti Money Laundering and Creditworthiness check. Once all the required conditions are met then the request proceeds into reference data set up and static data set up, which leads to the fulfilment of a new request. And the new client is now on-boarded on to the Banks platform for further trading activities.

The challenge the industry currently facing is in the areas of data compliance and data quality. Especially in the Investment Banking industry last few years has been a rough ride to manage with all the risk and compliance norms enforced by the Government and other key agencies like BCBS. This means the time has come for the Banks to up their ante for quick changes in the way they used to traditionally work. Banks need to be technologically advanced and have a game plan to be a tech-led service provider to stay in the game. And this is pretty evident in Investment Banks who are becoming more digitized and technology led (Wojcik et.al. 2016)

However, there is a certain gap when it comes to some of the core delivery processes. Where either due to the sheer size of the legacy data and accumulated systems or budget constraints, most of the Investment Banks are not in a position to bring in the change at the desired pace. This leaves some of the core processes as-is. In the sense that things are still done manually and in a rudimentary manner.

Keeping aside the technology challenge, when one considers the power of big data and leveraging analytics and insights and some of the Machine Learning techniques, one can help solve some of the basic yet key issues that Investment Banks face today as discussed in previous lines. These interventions will certainly help the Banks to optimize their workforce and streamline their process further. Creating a conducive environment for technology led innovations and solutions. The idea is to drive a point that Analytics & Insights not only works for a complex scenario but also in some simple basic but highly manual and rudimentary scenarios.

This paper attempts to discuss the process to onboard a new client and how through applying machine learning techniques and in particular decision tree one could effectively predict the transactions performed during the onboarding process is whether good or bad.

2.0 LITERATURE REVIEW

While there are a number of good articles and blogs that talk about prediction of financial fraud (Ngai et.al., 2011), fraudulent financial statements (Kirkos et.al., 2007), Bankruptcy prediction in banks (Kumar et.al., 2007), risk determination and management using predictive modelling (Gopinathan et.al., 2001) and many more in the context of a Bank. However, there is a gap when it comes to articles that showcase and explain the usage of Machine Learning and in particular decision-tree techniques to predict errors at a transaction level. The following are some high-level activities performed by an Investment Bank, Proprietary Trading, Merchant Banking, Asset Management, Advisory function etc. to name a few. The focus area of this paper is in the scope of trading business.

The data set used for building the decision tree model consists of 2,25,478 transactions performed by the delivery team for the last 6 months. There are 10 independent variables and the dependent variable is Returnflag 0 or 1. Where 0 is good and 1 is bad and needed correction/ re-work. The class distribution for 0 and 1 is 99.25 and 0.75 percent respectively. The imbalance in the data is likely to skew the decision and classify all the observations as 0 i.e. True Negative in this case. As the objective is to identify bad transactions from the good ones, certain techniques had to be deployed to deal with imbalanced data.

R provides some good packages to work with such imbalanced data sets. The following are some of the methods to deal with imbalanced data sets;

1) Undersampling
2) Oversampling
3) Synthetic data generation

2.1 Undersampling

This technique reduces the number of observations from majority class to make the data set balanced. This method is best to use when the data set is huge. Undersampling methods are of two types; Random and Informative

Random undersampling method randomly chooses observations from majority class which are eliminated until the data set gets balanced. Informative undersampling follows a pre-specified selection criterion to remove the observations from majority class (Ngai et.al., 2011), (Donoho & Tanner, 2010), (Liu et al, 2009)

2.2 Oversampling

This method works with minority class. It replicates the observations from minority class to balance the data. It is also known as upsampling. Similar to undersampling, this method also can be divided into two types: Random Oversampling and Informative Oversampling. Random oversampling balances the data by randomly oversampling the minority class. Informative oversampling uses a pre-specified criterion and synthetically generates minority class observations (Ngai et.al., 2011; Chawla et al., 2002; Deepa et al., 2010)

2.3 Synthetic data generation

In simple words, instead of replicating and adding the observations from the minority class, it overcomes imbalances by generating artificial data. It is also a type of oversampling technique. In regards to synthetic data generation, Synthetic Minority Oversampling Technique (SMOTE) is a powerful and widely used method. SMOTE algorithm creates artificial data based on feature space (rather than data space) similarities from minority samples. In other words, it generates a random set of minority class observations to shift the classifier learning bias towards minority class (Ngai et.al., 2011; Lundin et al, 2002; Ho et al., 1995).

Apart from the above, there are few more techniques like K-fold Cross-Validation, ensemble models etc that can be considered to overcome the challenge of imbalanced data (Kirkos et.al., 2007).

While one explores some of the above-mentioned techniques, it is important to understand how the value of the information is either extrapolated or degenerated can have an impact on the outcome of the prediction. One needs to find a fine balance while working with some of the above techniques.

The objective of this paper is to use some of the techniques of Machine Learning, in particular, classification analysis using decision tree algorithm to predict whether a transaction performed by an associate has some quality issues and hence 'bad or it is 'good' to be flown into the downstream systems. The ideal would be Straight Through Process without any manual touch point, but as highlighted above, this requires budget and resources from the Bank IT to bring in a systemic change. While the Banks have embarked on a path to enhance technology, this would be achieved over a period not less than two to three years' time.

The reason for selecting decision tree algorithm in particular for doing this classification analysis is due to its simplicity to understand and execute. And it works pretty well with both categorical and continuous data set.

3.0 METHODOLOGY

The dataset used for this study is the fulfilment process where the following key requests are handled; account opening, account closure, account maintenance and SSI set-up across different products and regions. The dataset had about 1600 odd records that have returned for re-work and in the EDA phase helped to understand and get some useful insights on specific areas to be focused.

3.1 Exploratory Data Analysis

The R script of EDA and final model is uploaded to GitHub. The same can be accessed through the link here. Below is a brief view of the data set is shown in Figure 1.

Product	Process	Region...Old	Maker	Type.of.activity	Request.Completed.Time1	TeamName	No.of.app.touched	Other_Sources	Actual_Vol	Returnflag
EQ	AO	APAC	dhaspiy	Cash EQ	5 am to 11 am	Equities - HK	3	xxCAR	1	0
FI	AO	APAC	hegdjyo	Agent (Including Domino Template)	5 am to 11 am	Sydney Static	2	xxCAR	1	0
FI	AO	APAC	hegdjyo	E / D Account (Including Domino Template)	5 am to 11 am	Sydney Static	2	xxCAR	1	0
FI	AO	APAC	sapkoud	E / D Account (Including Domino Template)	5 am to 11 am	Sydney Static	2	xxCAR	1	0
EQ	AM	APAC	thyacha	HK Govt	11 am to 4 pm	Equities - HK	1	Email	2	0

Fig. 1.Snapshot of the data set

There were 10 independent variables, 2 numerical and discrete and the rest are categorical. The dependent variable is 'Returnflag'- 0 and 1

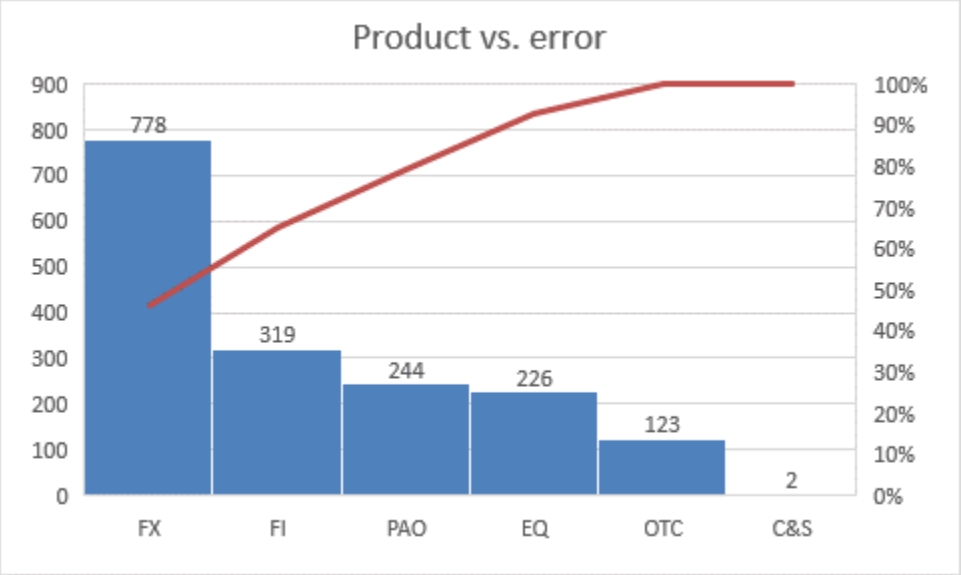

Fig. 2. No. of errors observed across products

It is observed that the errors in FX are higher than compared to other products, and 80% of the errors are observed in FX, FI &

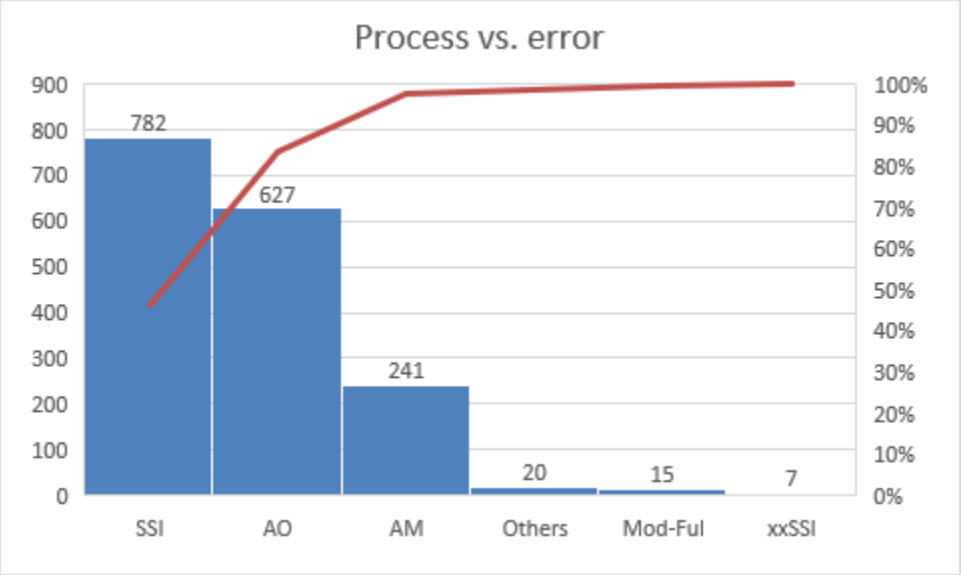

Fig 3. No. of errors observed in process

It is observed that the errors in SSI are higher than compared to other processes, and 80% of the errors are observed in SSI & AO processes.

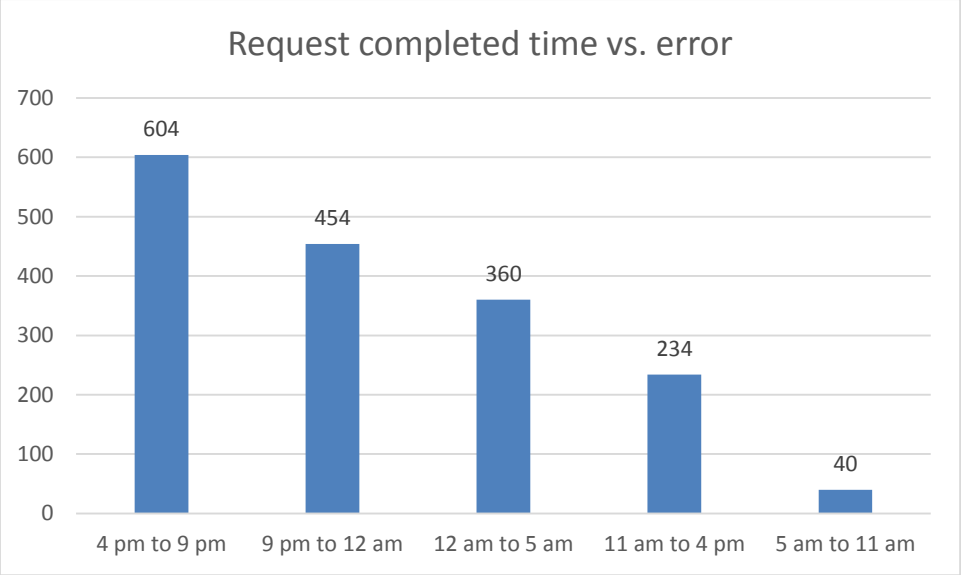

Fig 4. No. of errors observed across time slot

It is observed the errors are high for the transactions performed between 4 pm to 9 pm IST.

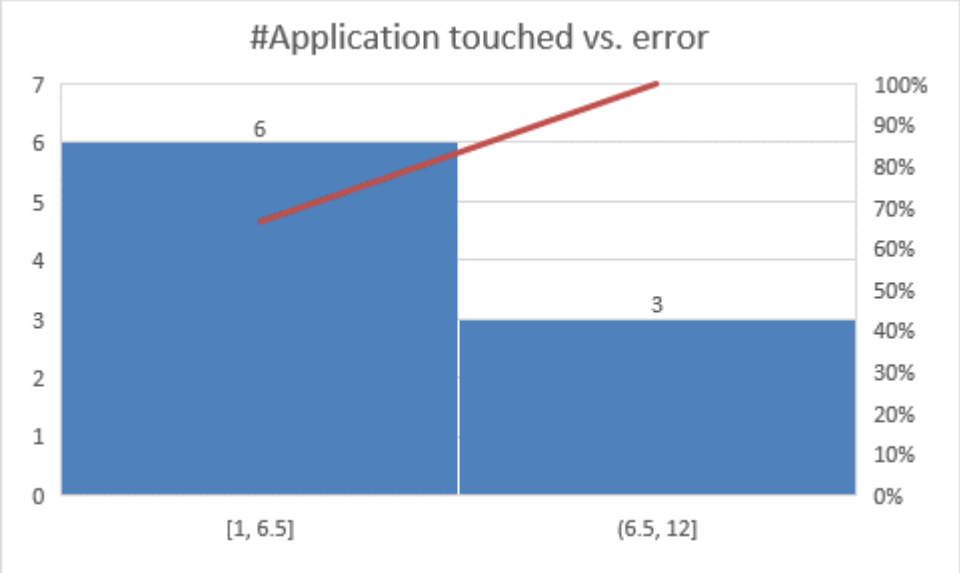

Fig 5. No. of applications touched vs. error

It is observed that 80% of the errors are for transactions that have the number of applications to be touched in the range of 1-7

3.2 Building the Decision Tree model with various sampling techniques

3.2.1 With as-is data

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 67117   302
         1    19   205

               Accuracy : 0.9953
                 95% CI : (0.9947, 0.9958)
    No Information Rate : 0.9925
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5588
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9997
            Specificity : 0.4043
         Pos Pred Value : 0.9955
         Neg Pred Value : 0.9152
             Prevalence : 0.9925
         Detection Rate : 0.9922
   Detection Prevalence : 0.9967
      Balanced Accuracy : 0.7020
```

Fig 6. Prediction summary of as-is data

3.2.2 With Oversampling

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 43451  4326
         1  3365 15993

               Accuracy : 0.8854
                 95% CI : (0.883, 0.8878)
    No Information Rate : 0.6973
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7249
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9281
            Specificity : 0.7871
         Pos Pred Value : 0.9095
         Neg Pred Value : 0.8262
             Prevalence : 0.6973
         Detection Rate : 0.6472
   Detection Prevalence : 0.7117
      Balanced Accuracy : 0.8576
```

Fig 7. Prediction summary for Oversampling

3.2.3 With Undersampling

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 13698  1767
         1  2224 12311

               Accuracy : 0.867
                 95% CI : (0.8631, 0.8708)
    No Information Rate : 0.5307
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7334
 Mcnemar's Test P-Value : 5.271e-13

            Sensitivity : 0.8603
            Specificity : 0.8745
         Pos Pred Value : 0.8857
         Neg Pred Value : 0.8470
             Prevalence : 0.5307
         Detection Rate : 0.4566
   Detection Prevalence : 0.5155
      Balanced Accuracy : 0.8674
```

Fig 8. Prediction summary of Undersampling

3.2.4 Both Over and Undersampling

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 13017  1992
         1  1925 13066

               Accuracy : 0.8694
                 95% CI : (0.8656, 0.8732)
    No Information Rate : 0.5019
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7389
 Mcnemar's Test P-Value : 0.2916

            Sensitivity : 0.8712
            Specificity : 0.8677
         Pos Pred Value : 0.8673
         Neg Pred Value : 0.8716
             Prevalence : 0.4981
         Detection Rate : 0.4339
   Detection Prevalence : 0.5003
      Balanced Accuracy : 0.8694
```

Fig 9. Prediction summary of both over and undersampling

3.2.5 With Synthetic Data Generation

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 9116 1468
         1 1369 9047

               Accuracy : 0.8649
                 95% CI : (0.8602, 0.8695)
    No Information Rate : 0.5007
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.7298
 Mcnemar's Test P-Value : 0.06578

            Sensitivity : 0.8694
            Specificity : 0.8604
         Pos Pred Value : 0.8613
         Neg Pred Value : 0.8686
             Prevalence : 0.4993
         Detection Rate : 0.4341
   Detection Prevalence : 0.5040
      Balanced Accuracy : 0.8649
```

Fig 10. Prediction summary of synthetic data

The key areas of focus for improvements are FI, FX and PAO, for SSI and AO processes. At the next level, it is suggested to further investigate the reason for high errors during the time slot of 4-9 pm IST.

		Model Result	
Sn No	sampling Method	Specificity/ Recall	Overall Accuracy
1	As-is	40%	99%
2	Oversampling	78%	88%
3	Undersampling	87%	86%
4	Both Over and undersampling	86%	86%
5	Synthetic data generation	86%	86%

Table 1: Summary of the model scores

4.0 CONCLUSION

Based the statistics of the prediction models across the four different sampling techniques, the numbers for undersampling looks better. Since we are dealing with imbalanced data, the key metric considered was ‘specificity’ or the ‘recall’ rate. This signifies the ability of the model to accurately identify bad transactions. The recall rate or the specificity score received for

undersampling was 87% and the overall accuracy of the model is 86%. Hence for the given data set undersampling technique is adopted for building the decision tree model.

5.0 SUMMARY

The objective of this paper was to explain machine learning using decision tree model for predicting errors or bad transactions at a transactional level in an Investment Bank scenario. Since the data collected was imbalanced, four different sampling techniques were tried, namely, oversampling, undersampling, both over and under sampling combined and synthetic data generation. Out of which undersampling gave the best recall rate of 87% and overall accuracy of the model is 86%. At the exploratory data analysis stage, we tried to gauge any significant patterns or trends in the data through graphs and visualization. Before building the model, to understand the significance of each of the independent variables information value technique was used, however except for one variable rest had a very weak or suspicious importance of the dependent variable. The model was built using all the 10 variables. Based on the model scores under various sampling techniques, undersampling provided the maximum accuracy and hence was used to build the final decision tree model.

REFERENCES

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Deepa, V. B., Thangaraj, P., & Chitra, S. (2010). Mining rare event classes in noisy EEG by oversampling techniques. In *Innovative Computing Technologies (ICICT), 2010 International Conference on* (pp. 1-6). IEEE.

Donoho, D. L., & Tanner, J. (2010). Precise undersampling theorems. Proceedings of the IEEE, 98(6), 913-924.

Gopinathan, K. M., Jost, A., Biafore, L. S., Ferguson, W. M., Lazarus, M. A., & Pathria, A. K. (2001). *U.S. Patent No. 6,330,546*. Washington, DC: U.S. Patent and Trademark Office.

Ho, T. K., & Baird, H. S. (1995). Evaluation of OCR accuracy using synthetic data. In Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval.

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert systems with applications*, *32*(4), 995-1003.

Kumar, P. R., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review. *European journal of operational research*, *180*(1), 1-28.

Liu, X. Y., Wu, J., & Zhou, Z. H. (2009). Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39(2), 539-550.

Lundin, E., Kvarnström, H., & Jonsson, E. (2002). A synthetic fraud data generation methodology. Information and Communications Security, 265-277.

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of the literature. *Decision Support Systems*, *50*(3), 559-569.

Wójcik, D., Knight, E., O'Neill, P., & Pažitka, V. (2016). *Investment banking since 2008: the geography of shrinkage and shift*. Financial Geography Working Paper. Available online at www. fingeo. net.

WEBLIOGRAPHY

https://www.cognizant.com/InsightsWhitepapers/Efficient-Client-Onboarding-The-Key-to-Empowering-Banks.pdf (Last accessed October 20th, 2017)

https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/ (Last accessed on October 27th 2017)

https://www.kdnuggets.com/2017/06/7-techniques-handle-imbalanced-data.html (Last accessed October 20th 2017)

http://crackmba.com/investment-banking-and-its-functions/ (Last accessed October 28th 2017)

GLOSSARY

1. BCBS – Basil Committee for Banking Security
2. EDA – Exploratory Data Analysis
3. IB – Investment Bank
4. IT – Information Technology
5. FI – Fixed Income
6. FX – Foreign Exchange
7. EQ – Equities
8. C&S – Credit and Swaps
9. PAO – Pxxxxxx Account Opening (Client specific, hence masked)
10. OTC – Over the Counter
11. SSI – Standard Settlement Instructions
12. AO – Account Opening
13. AM – Account Maintenance