**Bank Marketing Data Classification**

**Abstract**

This document sequentially applies a set of Data Science techniques to gain insights from the Direct Marketing campaign of a Portuguese Banking Institution. Data Science analysis of this data will benefit the business processes of the Banking and Financial Management Industry.

The Portuguese Bank that initiated the telemarketing campaign, (that provided the data examined by this document), contacted potential savings account depositors for a 5 year period, 2008 through 2013. This data therefore reflects the influence of the financial crisis of 2008. The original study gathered information on 150 different data categories, that covered information about clients, the bank’s products, social factors and economics. The original data was processed through data modelling with the objective of reducing the feature set. The data examined during that process was pre-July 2012 data. The result of the feature selection was a dataset of 22 of the starting 150 categories. This document examines the dataset of 22 marketing campaign metadata categories. A final set of 2 data categories, “Housing” and “Loan”, was determined as having the greatest effect on bank customers subscribing to “Term Deposit” accounts. Thereby, allowing for prediction of Term Deposit subscriptions by bank customers.

Another reason the examination of the telemarketing data of this Portuguese Bank is important is the controversy related to local marketing of banking products. National networks of large banks have encroached on the accessibility in local areas to small, locally-based banks. This expansion into the local territories of small banks has increased with legislation friendly to powerful, national banking enterprises. The results of this phenomena are the movement of bank deposit funds to large national banks, and the setting of prices for banking services on a non-local scale, the ubiquitousness of compatible ATM machines, and the availability of internet banking.

Thereby, not only is the availability of local banks affected by the expansion of national banks, the availability of banking products unique to local banks are also affected. Competition between banking institutions have moved from a local focus defined as the area of one city or one county, to a national focus defined as the area of one country. The large national banks in question are not limited to supplying banking products and services only, unlike previous local banks. The Horizontal Merger Guidelines of the Department of Justice and Federal Trade Commission, defines a market as a “product or group of products and a geographic area such that a hypothetical profit-maximizing firm, not subject to price regulation, that was the only present and future producer or seller of those products in that area likely would impose at least a ‘small but significant and nontransitory’ increase in price, assuming the terms of sale of all other products are held constant.”

In 1963, the Supreme Court held that the antitrust laws, and in particular section 7 of the Clayton Act (1914), applied to banking. In 1966, Congress reaffirmed the Supreme Court’s opinion regarding application of the Clayton Act via amending the Bank Merger Act of 1960, and the Bank Holding Company Act of 1956. In 1978, Congress further reaffirmed the Supreme Court’s opinion via passing the Change in Bank Control Act.

Modern geographic and product definitions used by banking institutions were examined by the Supreme Court during the Philadelphia National Bank case. The Court determined that because banking institutions were mostly local in scope, the local geographic area is relevant for analysis of competition in banking. The Supreme Court determined that “the cluster of products (various kinds of credit) and services (such as checking accounts and trust administration) denoted by the term ‘commercial banking’ composes a distinct line of commerce.” Therefore, local banking services and products are possibly free of effective competition from other banking institutions because of the locally distinctive characteristics of the services. Local banking institutions are also possibly exempted from competition by cost advantages, and customer preferences. Substitution of these services was often not possible on a local level.

Since 1963, antitrust courts have had to adapt to financial sector changes involving product line diversity, and access to local markets by national institutions. However, not until recently has evidence emerged of a shift to banking with services and products provided by national institutions. Antitrust Courts now have the statistical evidence needed for a redefinition of banking services and products. The evidence of changes in financial services includes research on the Survey of Consumer Finances (SCF) and the Surveys of Small Business Finances (SSBF) in 1993, 1998, and 2003.

The SSBF indicated that within the USA, small businesses obtain an average of two financial services from a local financial institution, and a local depository institution, that are primarily commercial banks. This is contrasted by small businesses obtaining only one banking service from a non-local institution. From 1989 to 1998, evidence of a shift to non-local services by individual households began to appear. During that time, consumers had began to rely on an increased amount of financial services, and the percentage of financial services obtained from one institution had decreased. As of 2003, most banking services and products were obtained by households and businesses from local banking institutions.

In 1998, 82 percent of all small business financial services were obtained from local banking institutions, with 94 percent of checking and savings services, and over 75 percent of financial management services. A survey conducted by the National Federation of Independent Businesses found that small businesses perceived local banking as a preferential option. Past SCF research indicated that households primarily relied on local banking institutions for banking transactions, certificates of deposits, and lines of credits. Yet households had started to rely on non-local services for alternate forms of banking. Even though the local areas in question had divergent deposit and loan rates, the suppliers of banking services remained local. In the early 1990s, higher loan interest rates, and significantly lower deposit interest rates were available via local banking.

The period from 2008 - 2013 represented an accelerated period of expansion of non-local, internet based banking options for individuals searching for banking services. During that period of time the services from local banks that were sought out by customers moved to national institutions. In order for new large national banking enterprises to sell banking products to previously local banking customers, an understanding of local factors of banking product selection is required. In 1994, only 1 percent of banking institutions did not have a branch in the local marketing area their customers lived in. By 2004, 18 percent of banking institutions did not have a local branch in order to respond to the individual needs of their customers.

Via Customer Segmentation with Data Science techniques, the previous local demands of banking services are conformable to corresponding segments of the population, determined by Customer Data characteristics. This segmentation is then usable by banking regulation institutions, and by businesses seeking to provide innovative banking services on a national scale. The effects of national banking services on all populations, national and local, are measureable with the causes of interest by individuals defined by data categories that are measurable with a national or local focus.

The Financial Services Modernization Act of 1999, has introduced complexity to the definition of banking service demand, and therefore the measurement of banking service marketing effectiveness and scope. As a result, the variety of banking services has grown to encompass the growing complexity of services defined as banking services.

Via the internet, banking service providers have expanded the range of services they have traditionally offered to customers. The expanded services now exist has separate business areas that provide deposit, loan, mortgage, credit, transaction card, vehicle loan, and business loan services. High risk short term loans, and investment brokerages, have become available with the same convenience of all other types of banking services. The origin corporations associated with new internet banking products have been obscured. Thereby, acceptance of banking product services has become independent of the enterprise providing the service.

Thereby, a customer-based focus of analysis of banking services via Data Science, allows for understanding of the possible effects of the concentration of a wide variety of banking resources into a small group of national enterprises. Divergent demographic and economic characteristics of consumers are now examinable independent of geographic area, in order to determine the likelihood of procurement of financial services.

**Synopsis**

This document utilizes Data Classification to examine a dataset related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The objective of the classification is to predict if the client will subscribe to a Term Deposit. “Data Classification” is the use of Machine Learning techniques to organize datasets into related sub-populations, not previous specified in the dataset. This can uncover hidden characteristics within data, and identify hidden categories that new data belongs within.

This document mainly utilizes the Data Science technique of “Data Classification” to examine a dataset related with direct marketing campaigns (telemarketing phone calls) of a Portuguese banking institution. The objective of the classification is to predict if the client will subscribe to a Term Deposit. “Data Classification” is the use of Machine Learning techniques to organize datasets into related sub-populations, not previous specified in the dataset. This can uncover hidden characteristics within data, and identify hidden categories that new data belongs within.

The Data Science techniques used within this research effort are Exploratory Data Analysis, Data Classification via K-means Clustering, Data Correlation Testing, Predictive Analytics, and Machine Learning with Cross-Validation.

The above forms of analysis are sequentially applied to refine the exploration of the Bank Marketing data, in order to determine if a given bank customer, (after the metadata that describes them has been recorded), will have a propensity to select a Term Deposit account as the form of account they are using to save their money. The most efficient method of deducing whether a customer will select a Term Deposit plan, after Exploratory Data Analysis, is Cross-Tabulation of the metadata categories to determine the distribution of previous Term Deposit acceptance. Via Cross-Tabulation, the categories of data with the least independent relationship to Term Deposits are selected for Classification according to the means of numeric variables. This form of Classification, K-means Clustering, creates easily visualized data clusters where future customer data are grouped within to facilitate Customer Segmentation for marketing purposes.

The next stage of Data Science processing of the Bank Marketing data involves Correlation Analysis of the explanatory data categories discovered via Cross-Tabulation and verified with K-Means Clustering. The Correlation test provides the ability to select a final set of data categories in the Bank Marketing data, that will allow for the greatest effectiveness, accuracy, and precision in selecting the form of Predictive Analytics to use for the final Machine Learning algorithm, (that will process future Bank Marketing data to determine affinity for subscribing to a Term Deposit). The Correlation test provides the status of the Alternate Hypothesis of data correlation, (whether it is possible to say the data is not in any way correlated), the Probability Value of a change in one data category causing a change in the second examined data category, and the 95% Confidence Interval of data independence.

After Correlation testing, the data variables within the Bank Marketing dataset that have indicated the greatest correlation to Term Deposit acceptance, are examined with two methods of Predictive Analytics to determine the Predictive Analytics algorithm to use for Cross-Validated Machine Learning. Predictive Analytics is a form of Data Science, where Probability Theory, (derived from the mathematical discoveries of Bayes, Kolmogorov. and Bernoulli), are applied to data categories to determine causality in data variables. Given the probability of data categories having an effect on other data categories, it is possible to deduce previous unknown information in datasets. The data that was hidden, before Predictive Analytics, is usable in a wide variety of business and government applications. Predictive Analytics examines data variables as Explanatory, or Independent, variables that possible have an effect on Response, or Dependent, variables. Usually several Explanatory Variables are processable via Predictive Analytics to discern the unknown state of one Response variable, however overfitting of Explanatory variables is possible, thereby leading to inaccurate prediction of the unknown state of the examined Response variable.

In choosing a Predictive Analytics algorithm for this project, high bias/low variance classifiers such as Naive Bayes, Linear Regression, Linear Discriminant Analysis and Logistic Regression were eliminated, considered the large amount of records in the Bank Marketing dataset. High bias suggest a wider variety of assumptions about the target data, than exists in this dataset. Low variance suggests small changes to the target as the training dataset changes. This isn’t the case with the categories of the dataset. Instead the classifiers considered for this analysis were low bias/high variance. Low bias presumes less assumptions about the target value. High variance presumes the target function will change greatly with slight changes in the training data. The low bias/high variance classifiers considered were Decision Trees, K-Nearest Neighbors, Support Vector Machines, and Random Forest.

Decision Trees were rejected because of inability to learn after initial processing, and possible overfitting of the data, therefore non-adaptable to new data. KNN was rejected because of the internal random number function causing irregular predictions, deemed not desirable for the goals of this project. Neural Networks were rejected because of inaccuracy with training datasets that do not have an extremely high dimensionality. Support Vector Machines were chosen for the trial stage of selecting a Predictive Algorithm for Machine Learning for high accuracy, resistance to overfitting, and the high dimensionality of the dataset. The Random Forest algorithm was chosen for the variety of classes, numeric/categorical, within the Bank Marketing dataset. Random Forest is also probability-based, thereby compensating for the distance-based scheme of Support Vector Machines.

**Working Directory, and Required Packages**

The RStudio IDE is used for interactive programming of the Data Science analysis, of the Bank Marketing data. In addition to the basic capabilities of the R programming language, several R language packages of pre-programmed functions are used for the analysis. These R packages include, “ggplot2”, “knitr”, “cluster”, “HSAUR”, “fpc”, “lattice”, “rpart”, “kernlab”, and “randomForest”. The Bank Marketing datasets are cleaned via use of “strings as factors” importing of the unorganized .csv files containing the data. The cleaned datasets are then stored as new .csv files, with new names designating that they are the cleaned version of the original dataset files.

The “knitr” table function, “kable()” is used for formatting the document’s tables. “ggplot2” bar graphs are used for the initial Exploratory Data Analysis. The base R function, “xtabs()” is used for Cross-Tabulation. Subsetting methods in R assist to create readability of subsequent clustering graphs. K-means Clustering and Plotting are used for Data Classification exploration. The “cor.test()” and “levelplot()” functions are used for correlation testing of the dataset’s explanatory variables. The “ksvm()” and “randomForest()” functions are used for Predictive Analytics. 100 Decision Trees are created for the Random Forest Machine Learning of the dataset, with a Cross-Validation scheme of 90% Training data, and 10% Testing data. Finally, the categorical variables in the dataset, that were converted to numeric variables for linear modelling within Predictive Analytics, are converted back to categorical variables, using the “lapply()” function. “lapply()” allows for complex looping through the dataset, without writing loop functions. The “unique()” function finds the category names, and the “unlist()” function accesses the output of the “lapply()” function for matching of converted numeric variables with original categorical variables.