**ISQA 8080 – Course Project Milestone 2**

CHAD CROWE, EL KERNS, AND JODI SANGSTER

Abstract (optional).

Overall: The project report should be written in a way that it is a) self-contained, i.e., has all information required to fully understand the entire document without having to consult external sources, b) captures all the activities that you conducted for the course project and follow the CRISP-DM cycle, and c) should be readable by non-experts, i.e., should provide explanation of the (statistical) concepts used.

1. Course Project Description and Business Understanding

The goal of this project is to build a model to predict churn or customer attrition for Intrado (formerly West Corporation), a global provider of communication and network infrastructure services. Acquiring and maintaining business relationships is necessary for companies to prosper. High rates of churn can affect profits and reduce company growth as it is more expensive to recruit new clients rather than retain existing ones. Churn is of especial concern in the communication industry as competition allows customers to transfer between service providers easily.

Data pertaining to Intrado’s customers has been collected both internally and purchased from external sources. This data can be used in conjunction with statistical machine learning methods to identify factors that influence a client’s propensity to churn. Identifying clients who have a high potential to churn allows for the implementation of targeted strategies to improve customer retention. We aim to produce a model that both predicts the probability of churn with a high degree of accuracy and can assign a prediction of churn at a threshold with a good trade-off between specificity and sensitivity.

1. Data Understanding

Data collected internally pertaining to specific accounts (table 1) and external firmographic data (tables 2 and 3) have been combined to create a robust data set that can be used for predictive modeling. Overall, the data set contains 47 variables of various types (date, categorical, ordinal, and numeric) with 24,649 observations. Additionally, churn is expressed as a binary categorical variable with a 7% churn rate overall in the data set used to build the model.

**Table 1.** Summary of variables and descriptive statistics from internal customer data.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Type** | **Missing values (%)** | **Min** | **Max** | **1st q** | **Median** | **Mean** | **3rd q** |
| Company creation date | Date | 0.02 | 1993 | 2018 | 2007 | 2011 | 2010 | 2014 |
| Total products | Int | 0 | 1 | 10 | 1 | 1 | 1.587 | 2 |
| Total transactions | Int | 0 | 1 | 1491 | 7 | 24 | 46.24 | 55 |
| Total revenue | Num | 0 | -781981 | 20015125 | 73 | 920 | 33389 | 6037 |
| Total usage (minutes) | Num | 0 | 0 | 861482918 | 363 | 6571 | 667517 | 52281 |
| Total accounts | Int | 0 | 1 | 129 | 1 | 1 | 8.0 | 1 |

The internal data (Table 1) contains primarily numeric data that includes the total number of products, transactions, revenue billed, usage in minutes, and accounts for each customer included. While total revenue and usage in minutes span a wide range of values, the remaining variables are discrete integers. In all instances, the distribution is strongly skewed with a higher frequency of values in the lower range of possible values for each variable. This is understandable as the businesses represented in the data set vary drastically in size and type as evidenced by marked differences in the number of employees and revenue (Table 2) for companies and by the variety of industries represented (Table 3).

**Table 2.** Summary of select numerical variables and descriptive statistics from external firmographic data.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Type** | **Missing values (%)** | **Min** | **Max** | **1st q** | **Median** | **Mean** | **3rd q** |
| Year started | Date | 32.4 | 1665 | 2017 | 1985 | 2000 | 1991 | 2009 |
| Square footage | Num | 24.7 | 0 | 11000000 | 0 | 0 | 18644 | 5816 |
| Revenue (usd) | Num | 44.1 | 1 | 226 b | 1 m | 14.1 m | 529.9 m | 96.35 m |
| Number of family members | Int | 24.7 | 0 | 61720 | 0 | 4 | 398.8 | 112 |
| Employees at location | Int | 46.5 | 1 | 223000 | 6 | 25 | 230.2 | 100 |
| Total employees | Int | 24.7 | 0 | 380300 | 3 | 29 | 1170 | 200 |
| Domestic ultimate employees | Int | 24.7 | 0 | 91800 | 0 | 6.0 | 183 | 50 |
| Domestic ultimate revenue | Num | 24.7 | 0 | 239 b | 0.11 m | 5.68 m | 1.8 b | 108.1 m |

The firmographic data (Table 2) is skewed similarly to the internal data with distributions weighted towards the lower range of possible values. This is likely due to marked difference in size and types of companies included in the data set. The majority of firmographic data has missing variables amounting to approximately 24.7% of the total number of observations. However, several variables include missing values for greater than 50% of the observations included in the internal data. Categorical variables (Table 3) will at times duplicate information found in other variables. This can be seen in revenue, presented as both a numerical variable (Table 2) and as a categorical variable (table 3). Additionally, variables such as those related to industry (string) are duplicated in the us 1987 sic 1 as a numeric code. Relationships between select variables are presented in Table 4.

**Table 3.** Summary of select categorical variables from external firmographic data.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Levels** | **Missing values (%)** |
| Business code | 4 | 0 |
| Country code | 138 | 0 |
| Location type | 3 | 24.7 |
| Bemfab marketability | 6 | 24.7 |
| Public private indicator | 2 | 24.7 |
| Owns rents code | 3 | 24.7 |
| Subsidiary indicator | 2 | 24.7 |
| Manufacturing indicator | 2 | 24.7 |
| Legal status code | 10 | 24.7 |
| Import export agent | 7 | 24.7 |
| Status code | 3 | 24.7 |
| Population code | 10 | 24.7 |
| Hierarchy code | 18 | 24.7 |
| Revenue range | 10 | 24.7 |
| Global ultimate indicator | 2 | 24.9 |
| Major industry category | 10 | 24.7 |
| Related industry | 3117 | 24.7 |
| Line of business | 826 | 24.7 |
| Us 1987 sic 1 | 826 | 24.7 |

1. Data Preparation

The internal and external firmographic data were merged into a single table based on the company number used in both tables. Specific variables such as company creation date were defined as a date; while, categorical variables containing numbers were explicitly defined as factors to avoid confusion later on. The following subsections highlight the data preparation steps taken for this project. It is important to note, this was not a linear process. Steps were repeated and continued as necessary until arriving at the final product.

**Missing values.** Missing values for factors were replaced with ‘missing’ to allow for ease of use when building the model. Certain variables in the firmographic data either include a code for data which is not available as is the case for import export agent code “g” or appear to have inserted zeros for missing data such as the case with the year the company was started. These were consolidated into one level.

**Variable simplification.** Categorical variables containing a large number of levels were assessed to determine if simplification was possible. Company creation date was restricted to a 4 digit year; while, us 1987 sic 1 was restricted to the first 4 digits to create a single code equivalent to the major and related industry variables. This still resulted in 826 levels.

**Multicolinearity.** Multicolinearity occurs when a predictor variable can be explained reasonably well by another predictor. All variables were assessed for multicollinearity with functions included in the statistical software used for this project. It was determined the number of transactions was strongly correlated (0.7) to the number of products a company purchased. This was addressed during model development.

**Variable creation.** A variable was created to combine the related variables, transactions and products, as a single numerical value, transactions per products used. This allowed for use of both variables in the model without the issues that may arise from collinearity. Additionally, a variable was added as an indicator for missing values. It was determined the 6090 (24.7 %) missing values common to the firmographic data could be attributed to the same companies. Thus, 1581 companies are missing observation for 22 out of 30 variables.

**Variable elimination.** Since the combined data set contains 47 variables, it was necessary to work towards eliminating variables that are likely not useful or not of high quality. It was arbitrarily determined that variables with more than 50% missing values and categorical variables contained within another variable (Table 4) and / or with more than 100 levels should be eliminated from consideration at this time. This simplified the data set by removing 18 variables; 8 with over 50% missing, 10 with more than 100 levels which were contained in other variables.

**Imputing numeric values.** Missing numerical values were estimated using functions within the statistical software used for this project (specifically the bagimpute method of the preprocess function in the caret package). It is important to note, this process took place after data is divided (as described in the following section) to avoid bias during the testing procedure.

**Table 4.** Categorical variable hierarchy and level number.

|  |  |
| --- | --- |
| **Variable** | **Levels** |
| Business code | 4 |
| Country code | 138 |
| State | 392 |
| City | 11,646 |
| Zip | 21,506 |
| Major\_industry\_category\_name | 10 |
| Line\_of\_business | 946 |
| Us\_1987\_sic\_1 | 3,167 |
| Related\_industries | 5,900 |

**Additional information.** The data set was not directly evaluated for outliers at this time as this was easier to assess once model development began. Normalizing the data eliminated issues caused by potential outliers. Final variable breakdown by churn status can be found in the appendix (Tables A1-A2).

1. Data Modeling, Model Building

As our primary goal for the model was to identify the customer characteristics that lead to churn and since this inherently a classification problem, we aim to use one variation of each main type of classification model: a simple classifier, regression, and trees. The specific model types we used were k-nearest neighbor (KNN), Least Absolute Shrinkage and Selection Operator (LASSO), and boosted classification trees (BCT). The following plan for model construction maximized both performance and trade-off between sensitivity and specificity.

Moving forward, the available training set was randomly split into a training (70%) and test (30%) sets. For each of the three model types (KNN, LASSO, and BCT), model parameters was selected using 5-fold cross validation (Figure 1) to reduce bias and provide a more realistic estimate of model performance. The actual parameters being selected during cross validation varied between model types. For KNN, it was the number of neighbors used (k). For LASSO, this was the shrinkage parameter (λ). For BCT this was the number of trees (B), the shrinkage parameter (λ), and the tree depth (d). LASSO and KNN used pre-process function to center, scale, and impute missing numeric variables (using bagimpute) within each training fold as each of these model types perform better and offer increased interpretability when data are normalized and require that there are no missing values in the variables included in order to predict all cases. BCT was also run for each of the sets of models with imputed, but not standardized numeric variables as it is not required for good performance for this type of model.

The biggest challenge to model performance with this data set falls with the imbalance of data, approximately 92% of customers remain with the company while 8% have been lost due to churn and the correlation between total transactions and total products. To account for these problems, the 5-fold cross validation selection of parameters for each model type was repeated 6 times. Three models for each type was built using under-sampling techniques to create a more balanced data set. This technique restricts the majority class, customers who have not churned, from use during model building. Under-sampling may help to improve model sensitivity and performance in some cases. The remaining three models used the entire training data set for model construction. Of the three models built with under-sampling and three without, one included total transactions, one included total products, and the final one included a combined variable indicating avg. transactions/product (Figure 1).

Figure 1. Schematic for proposed models used to predict churn.

Selection in the cross-validation process was set to optimize for the receiver operating characteristic (ROC) curve allowing for a focus on sensitivity as there was a high degree of specificity due to the imbalance in data. Final models of each type was tried on the test set and the type leading to the greatest ROC was selected. The threshold used to make predictions on the company provided test set was determined based on what value lead to the highest sensitivity without requiring > 20% of the overall sample to be predicted positive.

4.1 K-Nearest Neighbor (KNN)  
Why did our group consider a K-Nearest-Neighbors model?

Our grouped selected a K-Nearest-Neighbors (KNN) model type for data training. KNN is a good model candidate because of its unique learning style. Its style classifies based on nearby data. This style sets it apart from other models that learn by separating data or finding patterns. This gives our group a range of model types. The hope is that each model type will learn different aspects of the target variable. Our group can then combine the data learned from each facet into a single, better model.

Which model were trained?

The model building section specifies three data combinations for model training. The three combinations concern the correlated variables total transactions and products. Our group decided to include variations of these variables into each model type.

1. total transactions without total products

2. total products without total transactions

3. total transactions / total products

The group wanted to include under-sampling in our models. Training included either under-sampling or no sampling for each model variation.Following suite with this methodology, we created six KNN models.

How did you train your models?

Model training included cross-validation. The data was also split into a 70-30 split. We evaluated the trained model using the test data. We report the performance via ROC and the confusion matrix on the test set.

Further training includes hyperparameter training. KNN is a simpler model. The focal parameter of hyperparameter training is k. K refers to the number of neighbors considered when classifying each record. We trained models with K values from 1 to 20. The preferred K value is the one which produces the highest ROC with cross-validation. Generally, we would prefer a K value greater than 1. This introduces model bias, but also creates less variability. By using K-values greater than 1, we can limit the variability. This allows us to consider a greater range of data when classifying with KNN.

How did you evaluate your models?

Why do you think your approach is appropriate for the given problem?

Which parameters did you use, and why? (this is very important with respect to replicability of a given study/project)

How did you select the parameters for your models?

Are there any particular assumptions that you used for model building (assumptions about data distributions, etc.)?

4.2 Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO was chosen as the regression model for this dataset for a few key reasons. LASSO differs from ordinary least squares in that the use a penalty parameter λ. This penalty parameter performs what is known as regularization/shrinkage which means that the larger the λ, the closer the beta-coefficients are shrunk to 0. This helps to avoid overfitting and works, unlike ordinary least squares, when your sample size is equal to or less than your number of variables. This datasets ratio of explanatory variables to sample size. It also handles multicollinearity better. Further, as the algorithm used largely depnds on unit variance thus the need to standardize the numeric variables, does not require the assumptions about the data that ordinary least squares does. LASSO was chosen over Ridge (another regularization method for regression) as LASSO is much more likely to set coefficients to exactly 0 and thus performs variable selection.

The LASSO models were created using the glmnet model in the caret package in R using 5-fold cross validation. The only parameter in need of optimization for this model type was λ. Optimization was completed by starting with a grid of length 10 containing values from 0.01-1 and then zeroing in on the range that the peak ROC seemed to fall within. The final value of lambda selected for each model was the maximum within one standard error of the one producing the maximum ROC. Once the models were optimized, performance was assessed using the hold-out test set taken from the original dataset provided. AUC for each model are presented in Table 6. Due to correlations found between the variables, total products and total transactions, 3 separate data sets were used to build the models. Dataset 1 contains the total products but not total transactions. Dataset 2 contains total transactions. Dataset 3 contains a composite variable created from the two correlated predictors, transactions per products. Additionally, each data set was evaluated both with and without undersampling.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 6: LASSO Model Results | | | |
| Variable Included | Sampling Type | Final Parmaeters | AUC |
| Transactions | None | λ = 0.0036 | 0.7014 |
| Transactions | Under | λ = 0.0215 | 0.6945 |
| Products | None | λ = 0.0013 | 0.6679 |
| Products | Under | λ = 0.0215 | 0.6610 |
| Transactions/Product | None | λ = 0.0013 | 0.7130 |
| Transactions/Product | Under | λ = 0.0013 | 0.7110 |

Models with the composite variable Transactions/Product performed better than those containing one or the other. Additionally, models with no sampling performed better than those with undersampling. The best performing model was the one containing the composite variable that used no sampling. The variables of importance in that model (ones with non-zero coefficients [for at least one level for factor variables]) are listed below:

* total\_accounts
* Business\_Code
* Subsidiary\_Indicator
* Legal\_Status\_Code
* Currency\_Code
* Population\_Code
* Hierarchy\_Code
* Revenue\_US\_Dollars\_
* Number\_of\_Family\_Members
* Major\_Industry\_Category\_Name
* TransactionsPerProduct

4.3 Boosted Classification Tree

A tree-based method was selected for use in this project since these techniques provide high-performing models which are easy to interpret. This type of predictive model relies on segmenting the predictors into simplified regions. We can use training data to define these regions and use the definitions to predict outcomes for new or unknown observations. To improve the performance of our model, we have chosen to use boosting. Boosting is an approach where trees are grown sequentially using information from the previous tree. Each subsequent generation of tree attempts to improve the model by updating the residuals, difference between observed and predicted values, produced by the previous generation of trees.

Boosted classification trees were created using the xgbTree model in the caret package in R using 5-fold cross validation. This particular model has 7 parameters that necessitate optimization (Table 7). Optimization was completed using a 5 step processs modified from (XXXX). A wide range of values was used to determine an initial value to use for the shrinkage parameter (eta) and to narrow the range of values used for the number of iterations (nrounds). These values were used to determine optimum values for the maximum interaction depth (max\_depth) and minimum child weight (min\_child\_weight). Next, values used for column and row sampling were determined followed by optimization of the gamma parameter. Finally, values for both nrounds and eta were reassessed to complete the process. Parameter values were selected that provided the highest ROC for the range of values tested.

**Table 7.** Hyper parameters used to create boosted decision tree models.

|  |  |  |
| --- | --- | --- |
| **Parameter Name** | **Description** | **Optimization Range** |
| nrounds | Iterations or number of tress used to learn | 10 - 1000 |
| max\_depth | Interaction depth controls the number of splits at each node | 3 - 9 |
| eta | Shrinkage parameter which controls the learning rate of the model | 0.01 – 0.1 |
| gamma | Minimum loss reduction required to further partition a node | 0 - 1 |
| colsample\_bytree | Fraction of columns used to subsample for each new depth level in a tree | 0 - 1 |
| min\_child\_weight | Minimum sum of instance weights needed to initiate further partitioning at a node | 1 - 3 |
| subsample | Subsample ratio of training data used to make each tree | 0 - 1 |

Once the models were optimized, performance was assessed using the hold-out test set taken from the original dataset provided. Accuracy, sensitivity, specificity, and AUC for each model are presented in Table 8. Due to correlations found between the variables, total products and total transactions, 3 separate data sets were used to build the models. Dataset 1 contains the total products but not total transactions. Dataset 2 contains total transactions. Dataset 3 contains a composite variable created from the two correlated predictors, transactions per products. Additionally, each data set was evaluated both with and without standardization and undersampling.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Criteria** | **Basic Model** | **With standardization** | **Undersampling** | **Undersampling with standardization** |
| Dataset 1 | Accuracy | 0.931 | 0.932 | 0.657 | 0.640 |
| Sensitivity | 0.033 | 0.046 | 0.664 | 0.678 |
| Specificity | 0.999 | 0.999 | 0.656 | 0.637 |
| AUC | 0.727 | 0.725 | 0.716 | 0.713 |
| Dataset 2 | Accuracy | 0.932 | 0.932 | 0.640 | 0.687 |
| Sensitivity | 0.048 | 0.054 | 0.678 | 0.649 |
| Specificity | 0.999 | 0.998 | 0.637 | 0.689 |
| AUC | 0.762 | 0.760 | 0.713 | 0.730 |
| Dataset 3 | Accuracy | 0.932 | 0.931 | 0.691 | 0.699 |
| Sensitivity | 0.062 | 0.048 | 0.651 | 0.653 |
| Specificity | 0.997 | 0.997 | 0.694 | 0.703 |
| AUC | 0.759 | 0.762 | 0.732 | 0.745 |

All models created without undersampling had comparable high accuracy and specificity with low sensitivity. However, the AUC was better for the models created using datasets 2 and 3. The models created using undersampling improved the sensitivity of the models as expected; however, both accuracy and AUC were sacrificed. Since the models using the predictor, total transactions, and the composite variable, transactions per product, performed similarly, it was decided that dataset 2 with only total transactions would be used as this required less data manipulation for future use. Additionally, standardization of numerical variables had little effect on the model performance. Thus, the basic model was used going forward. Optimized parameters for this model are presented in Table 9.

|  |  |
| --- | --- |
| **Parameter Name** | **Optimized Value** |
| nrounds | 140 |
| max\_depth | 7 |
| eta | 0.05 |
| gamma | 0.05 |
| colsample\_bytree | 0.5 |
| min\_child\_weight | 3 |
| subsample | 1 |

1. Evaluation, Recommendation, and Conclusion

How does the performance of the different models compare against each other?

Which evaluation metrics did you use, and why?

How do you interpret the results? What do they mean for the given project?

What are your recommendations for the client, and why? Which model did you use to create the predictions in your submission file?

What is your overall conclusion for the project and the data?

1. Presentations (Written and Oral)

Make sure to check for spelling and grammar.

Figures should be readable and explained /referred to in the report.

Put the main results (and summary tables/figures) in the main text, extended results in the Appendix (if applicable).

Powerpoint: prepare a 10-15 minute presentation of:

* Business and Data Understanding (brief)
* Your data preparation approach (e.g., filtering, outliers, sampling, etc.)
* The models that you used, and their parameters
* The evaluation of the different models
* Selection of best model
* Recommendation/Conclusion

1. REFERENCES (optional)

Include references where appropriate.

# Appendix

**Table A1. Comparison of continuous numerical variables based on whether customers remained with or churned from Intrado. Statistically significant differences (p < 0.05) are denoted in red type.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Retained** | **Churned** | **p-value** |
| **total\_products (mean (SD))** | 1.65 (1.04) | 1.51 (0.84) | <0.001 |
| **total\_transactions (mean (SD))** | 54.42 (79.58) | 31.64 (43.32) | <0.001 |
| **total\_revenue (mean (SD))** | 40820.72 (315008.87) | 17250.98 (129800.68) | 0.003 |
| **total\_usage (mean (SD))** | 824862.87 (11559048.81) | 362439.69 (3792051.04) | 0.111 |
| **total\_accounts (mean (SD))** | 8.31 (21.64) | 7.08 (21.27) | 0.029 |
| **Year\_Started (mean (SD))** | 1991.26 (26.79) | 1991.60 (26.41) | 0.624 |
| **Square\_Footage (mean (SD))** | 17896.91 (149273.85) | 23169.27 (89147.98) | 0.164 |
| **Revenue\_\_US\_Dollars\_ (mean (SD))** | 444038030.94 (3509402149.11) | 244411553.14 (1363459914.47) | 0.024 |
| **Number\_of\_Family\_Members (mean (SD))** | 387.84 (1948.40) | 232.80 (1126.03) | 0.002 |
| **Employees\_Here (mean (SD))** | 261.19 (2380.28) | 248.58 (2363.14) | 0.839 |
| **Employee\_Count\_Total (mean (SD))** | 1140.18 (9144.51) | 689.74 (3821.38) | 0.05 |
| **Domestic\_Ultimate\_Employee\_Count (mean (SD))** | 171.72 (1616.16) | 210.14 (2413.23) | 0.384 |
| **Domestic\_Ultimate\_Revenue (mean (SD))** | 1743638088.75 (10859344954.97) | 1364967026.52 (9659392035.25) | 0.177 |
| **YearCreated (mean (SD))** | 2010.52 (4.76) | 2010.26 (4.71) | 0.033 |
| **PropMissing (mean (SD))** | 0.08 (0.19) | 0.09 (0.21) | 0.722 |

**Table A2. Comparison of categorical variables based on whether customers remained with or left (churned) from Intrado.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Retained** | **Churned** | **p-value** |
| n | 18537 | 1603 |  |
| churned = 1 (%) | 0 ( 0.0) | 1603 (100.0) | <0.001 |
| **Business\_Code (%)** |  |  | <0.001 |
| APAC | 3357 (18.1) | 183 ( 11.4) |  |
| CANADA | 228 ( 1.2) | 34 ( 2.1) |  |
| EMEA | 6219 (33.5) | 367 ( 22.9) |  |
| USA | 8733 (47.1) | 1019 ( 63.6) |  |
| **Location\_Type (%)** |  |  | 0.012 |
| Branch | 1670 ( 9.0) | 110 ( 6.9) |  |
| HQ | 7105 (38.3) | 632 ( 39.4) |  |
| Single Location | 8325 (44.9) | 717 ( 44.7) |  |
| NA | 1437 ( 7.8) | 144 ( 9.0) |  |
| **Public\_Private\_Indicator (%)** |  |  | 0.091 |
| N | 16757 (90.4) | 1437 ( 89.6) |  |
| Y | 343 ( 1.9) | 22 ( 1.4) |  |
| NA | 1437 ( 7.8) | 144 ( 9.0) |  |
| **Owns\_Rents\_Code (%)** |  |  | <0.001 |
| 0 | 12672 (68.4) | 939 ( 58.6) |  |
| 1 | 1418 ( 7.6) | 171 ( 10.7) |  |
| 2 | 3010 (16.2) | 349 ( 21.8) |  |
| NA | 1437 ( 7.8) | 144 ( 9.0) |  |
| **Subsidiary\_Indicator (%)** |  |  | 0.03 |
| 0 | 11637 (62.8) | 956 ( 59.6) |  |
| 3 | 5463 (29.5) | 503 ( 31.4) |  |
| NA | 1437 ( 7.8) | 144 ( 9.0) |  |
| **Manufacturing\_Indicator (%)** |  |  | <0.001 |
| 0 | 10018 (54.0) | 677 ( 42.2) |  |
| 1 | 7082 (38.2) | 782 ( 48.8) |  |
| NA | 1437 ( 7.8) | 144 ( 9.0) |  |
| **Legal\_Status\_Code (%)** |  |  | 0.005 |
| 0 | 3656 (19.7) | 266 ( 16.6) |  |
| 3 | 11524 (62.2) | 1059 ( 66.1) |  |
| 8 | 32 ( 0.2) | 5 ( 0.3) |  |
| 12 | 1074 ( 5.8) | 75 ( 4.7) |  |
| 13 | 279 ( 1.5) | 16 ( 1.0) |  |
| 50 | 22 ( 0.1) | 0 ( 0.0) |  |
| 100 | 15 ( 0.1) | 1 ( 0.1) |  |
| 101 | 100 ( 0.5) | 6 ( 0.4) |  |
| 118 | 10 ( 0.1) | 0 ( 0.0) |  |
| 120 | 388 ( 2.1) | 31 ( 1.9) |  |
| NA | 1437 ( 7.8) | 144 ( 9.0) |  |
| **Currency\_Code (%)** |  |  | <0.001 |
| 10 | 17 ( 0.1) | 1 ( 0.1) |  |
| 20 | 7238 (39.0) | 838 ( 52.3) |  |
| 30 | 21 ( 0.1) | 4 ( 0.2) |  |
| 50 | 544 ( 2.9) | 36 ( 2.2) |  |
| 105 | 2 ( 0.0) | 0 ( 0.0) |  |
| 110 | 107 ( 0.6) | 21 ( 1.3) |  |
| 120 | 48 ( 0.3) | 7 ( 0.4) |  |
| 160 | 1977 (10.7) | 73 ( 4.6) |  |
| 190 | 1 ( 0.0) | 0 ( 0.0) |  |
| 220 | 94 ( 0.5) | 5 ( 0.3) |  |
| 230 | 258 ( 1.4) | 6 ( 0.4) |  |
| 250 | 11 ( 0.1) | 0 ( 0.0) |  |
| 280 | 339 ( 1.8) | 7 ( 0.4) |  |
| 290 | 1 ( 0.0) | 0 ( 0.0) |  |
| 310 | 4 ( 0.0) | 1 ( 0.1) |  |
| 320 | 31 ( 0.2) | 0 ( 0.0) |  |
| 330 | 41 ( 0.2) | 7 ( 0.4) |  |
| 370 | 125 ( 0.7) | 6 ( 0.4) |  |
| 380 | 56 ( 0.3) | 4 ( 0.2) |  |
| 390 | 3 ( 0.0) | 0 ( 0.0) |  |
| 400 | 44 ( 0.2) | 4 ( 0.2) |  |
| 430 | 1 ( 0.0) | 0 ( 0.0) |  |
| 440 | 1 ( 0.0) | 0 ( 0.0) |  |
| 500 | 7 ( 0.0) | 0 ( 0.0) |  |
| 560 | 1 ( 0.0) | 0 ( 0.0) |  |
| 580 | 22 ( 0.1) | 0 ( 0.0) |  |
| 620 | 57 ( 0.3) | 7 ( 0.4) |  |
| 790 | 15 ( 0.1) | 1 ( 0.1) |  |
| 800 | 1 ( 0.0) | 0 ( 0.0) |  |
| 820 | 14 ( 0.1) | 0 ( 0.0) |  |
| 970 | 116 ( 0.6) | 4 ( 0.2) |  |
| 1010 | 1 ( 0.0) | 0 ( 0.0) |  |
| 1060 | 5 ( 0.0) | 0 ( 0.0) |  |
| 1070 | 1 ( 0.0) | 0 ( 0.0) |  |
| 2010 | 4 ( 0.0) | 1 ( 0.1) |  |
| 2020 | 24 ( 0.1) | 4 ( 0.2) |  |
| 2080 | 2 ( 0.0) | 0 ( 0.0) |  |
| 3000 | 157 ( 0.8) | 11 ( 0.7) |  |
| 3030 | 9 ( 0.0) | 1 ( 0.1) |  |
| 3040 | 35 ( 0.2) | 5 ( 0.3) |  |
| 3060 | 3 ( 0.0) | 0 ( 0.0) |  |
| 4040 | 0 ( 0.0) | 1 ( 0.1) |  |
| 5000 | 1 ( 0.0) | 0 ( 0.0) |  |
| 5010 | 7 ( 0.0) | 0 ( 0.0) |  |
| 5040 | 9 ( 0.0) | 0 ( 0.0) |  |
| 5060 | 2 ( 0.0) | 0 ( 0.0) |  |
| 5080 | 1779 ( 9.6) | 140 ( 8.7) |  |
| 5090 | 1 ( 0.0) | 0 ( 0.0) |  |
| 6040 | 1 ( 0.0) | 0 ( 0.0) |  |
| 6050 | 1 ( 0.0) | 0 ( 0.0) |  |
| 6060 | 1 ( 0.0) | 0 ( 0.0) |  |
| 6600 | 1 ( 0.0) | 0 ( 0.0) |  |
| 6800 | 12 ( 0.1) | 2 ( 0.1) |  |
| 7200 | 3 ( 0.0) | 0 ( 0.0) |  |
| 7500 | 41 ( 0.2) | 3 ( 0.2) |  |
| 9100 | 26 ( 0.1) | 2 ( 0.1) |  |
| 9300 | 25 ( 0.1) | 3 ( 0.2) |  |
| 9410 | 76 ( 0.4) | 10 ( 0.6) |  |
| 9440 | 15 ( 0.1) | 1 ( 0.1) |  |
| 9450 | 20 ( 0.1) | 2 ( 0.1) |  |
| NA | 5078 (27.4) | 385 ( 24.0) |  |
| **Status\_Code (%)** |  |  | 0.012 |
| 0 | 8325 (44.9) | 717 ( 44.7) |  |
| 1 | 7105 (38.3) | 632 ( 39.4) |  |
| 2 | 1670 ( 9.0) | 110 ( 6.9) |  |
| NA | 1437 ( 7.8) | 144 ( 9.0) |  |
| **Population\_Code (%)** |  |  | <0.001 |
| 0 | 8836 (47.7) | 520 ( 32.4) |  |
| 1 | 2 ( 0.0) | 0 ( 0.0) |  |
| 2 | 3 ( 0.0) | 0 ( 0.0) |  |
| 3 | 14 ( 0.1) | 2 ( 0.1) |  |
| 4 | 102 ( 0.6) | 13 ( 0.8) |  |
| 5 | 222 ( 1.2) | 26 ( 1.6) |  |
| 6 | 349 ( 1.9) | 39 ( 2.4) |  |
| 7 | 757 ( 4.1) | 78 ( 4.9) |  |
| 8 | 1203 ( 6.5) | 114 ( 7.1) |  |
| 9 | 5612 (30.3) | 667 ( 41.6) |  |
| NA | 1437 ( 7.8) | 144 ( 9.0) |  |
| **Hierarchy\_Code (%)** |  |  | 0.008 |
| 0 | 6648 (35.9) | 548 ( 34.2) |  |
| 1 | 4236 (22.9) | 357 ( 22.3) |  |
| 2 | 2682 (14.5) | 225 ( 14.0) |  |
| 3 | 1522 ( 8.2) | 134 ( 8.4) |  |
| 4 | 839 ( 4.5) | 114 ( 7.1) |  |
| 5 | 515 ( 2.8) | 37 ( 2.3) |  |
| 6 | 252 ( 1.4) | 19 ( 1.2) |  |
| 7 | 151 ( 0.8) | 12 ( 0.7) |  |
| 8 | 94 ( 0.5) | 4 ( 0.2) |  |
| 9 | 53 ( 0.3) | 4 ( 0.2) |  |
| 10 | 33 ( 0.2) | 1 ( 0.1) |  |
| 11 | 22 ( 0.1) | 2 ( 0.1) |  |
| 12 | 23 ( 0.1) | 1 ( 0.1) |  |
| 13 | 21 ( 0.1) | 0 ( 0.0) |  |
| 14 | 3 ( 0.0) | 0 ( 0.0) |  |
| 15 | 3 ( 0.0) | 0 ( 0.0) |  |
| 16 | 2 ( 0.0) | 1 ( 0.1) |  |
| 19 | 1 ( 0.0) | 0 ( 0.0) |  |
| NA | 1437 ( 7.8) | 144 ( 9.0) |  |
| **Global\_Ultimate\_Indicator (%)** |  |  | 0.2 |
| N | 13530 (73.0) | 1148 ( 71.6) |  |
| Y | 3526 (19.0) | 307 ( 19.2) |  |
| NA | 1481 ( 8.0) | 148 ( 9.2) |  |
| **Major\_Industry\_Category (%)** |  |  | <0.001 |
| Agriculture, Forestry, Fishing | 68 ( 0.4) | 3 ( 0.2) |  |
| Construction | 320 ( 1.7) | 23 ( 1.4) |  |
| Finance, Insurance, Real Estate | 2357 (12.7) | 169 ( 10.5) |  |
| Manufacturing | 2577 (13.9) | 260 ( 16.2) |  |
| Mining | 117 ( 0.6) | 5 ( 0.3) |  |
| Public Administration | 864 ( 4.7) | 60 ( 3.7) |  |
| Retail Trade | 505 ( 2.7) | 46 ( 2.9) |  |
| Services | 8054 (43.4) | 683 ( 42.6) |  |
| Transportation & Public Utilities | 796 ( 4.3) | 48 ( 3.0) |  |
| Wholesale Trade | 1442 ( 7.8) | 162 ( 10.1) |  |
| NA | 1437 ( 7.8) | 144 ( 9.0) |  |