De-anonymization of Insurance Applicants' Sensitive Information

This research report was requested by the US National Association of Insurance Commissioners as part of their efforts to improve clients privacy on sensitive information.

The researchers in charge were:

Rosalie Dolor

Maxim Castañeda

Jay Lee

Team 3

SUMMARY

Over the years, as the life insurance industry has expanded, the need for clear regulatory laws has grown as well, originating entities like the National Association of Insurance Commissioners (NAIC). Through the NAIC, insurance regulators establish national standards and best practices, conduct peer reviews and coordinate their regulatory oversight to better protect the interests of consumers while ensuring a strong, viable insurance marketplace. American insurance regulators also take part in the International Association of Insurance Supervisors (IAIS) along with the NAIC by participating in all its major standard-setting initiatives, including working with fellow regulators from around the world to better supervise cross-border insurers, identifying systemic risk in the insurance sector and creating international best practices.

The NAIC is not the industry regulator itself, however it does work as intermediary between states insurance commissioners and federal instances. As pointed in their general mission, the NAIC seeks to: 1. Protect the public interest, 2.Promote competitive markets, 3.Facilitate fair and equitable treatment of insurance consumers, 4. Promote the reliability, solvency and financial solidity of insurance institutions, and 5. Support and improve state regulation of insurance. Hence, this research is a collaboration to enhance insurance companies' best practices by ensuring the clients' privacy rights in the information gathering process.

Prudential, one of the largest issuers of life insurance in the USA, held a Kaggle competition to predict the risk level of its applicants. The researchers used the data provided in the competition to build models capable of predicting applicants' sensitive attributes. The dataset contains 128 attributes (or variables) of 59,381 applicants. It contains information that applicants provided to Prudential when applying to the company's products.

The researchers used data mining methods through a series of steps to de-anonymize sensitive attributes from the given dataset. The process included: identification of sensitive variables, building of predictive models for de-anonymization, and exclusion of variables related to the sensitive variables. Given the anonymized dataset, the researchers checked if the removal of the sensitive variables affected risk level prediction. The flowchart in Figure 1 summarizes the whole methodology process that is followed in this research.

Having applied the de-anonymization process, it was found out that dropping the identified sensitive variables (and the important variables related to them) did not *significantly* affect the risk level prediction. However, assumptions made by the researchers regarding some of the attributes should be checked with Prudential. Also, the performance metrics that was used for evaluating the predictive models should be discussed with NAIC. Finally, it is recommended to repeat the process when new sensitive attributes are identified and then risk-level prediction should be reevaluated accordingly.

Problem Description, Business Goal and Data Mining Goal

The NAIC is concerned that insurance companies require some sensitive information from the applicants like their family medical records as part of the insurance application process. Such requirements are considered a bad practice from insurance companies whose clients might feel offended, and then discouraged to continue with the application process. Using a dataset from Prudential, NAIC intends to identify the most sensitive attributes from the life insurance application form and eliminate them from the questionnaire.

Following NAIC's goal, the data mining task is about building predictive models for the identified sensitive attributes of the applicants. The selection of sensitive variables was driven by reasoning from the customers point of view, deducing which information the applicants would feel uncomfortable to give away when applying for an insurance. Without considering time and financial constraints in the analysis, a good way to identify sensitive information for the applicants would be to conduct a survey. Without the possibility of conducting such survey and considering these constraints, the researchers took the freedom to rationally decide which variables are deemed sensitive, ranking them as described next: 1) family history (Appendix C), 2) employment and income information, 3) historical medical records and 4) insurance history.

Analyzing the available attributes, the team made the conclusion that insurance history and historical medical records could be easily gathered from the national healthcare system, and are necessary for the classification of the applicant's risk level. Nevertheless, family history and employment (income) information were then considered as the most sensitive variables. For *Family History*, some attributes are related to death of family members from specific kind of diseases. The researchers believe that this death-related information will make the applicants less attractive to the insurance companies. It might lower the applicants' chance to qualify for a life insurance or it could also be that they would have to pay for more expensive level of insurance because of circumstances (relatives dying from chronic diseases) that are completely beyond their control. Hence, applicants might be reluctant to share this kind of personal information.

Data Description

The dataset containing attributes of life insurance applicants was taken from a Kaggle competition arranged by Prudential, a US insurance company. After cleaning the dataset, the final version, which is used for building the models, contains 59,381 rows (each row is an insurance applicant) and over 900 columns because categorical attributes are turned into dummy variables. The output variables in this study are *Family_Hist_1* (categorical variable with three levels pertaining to whether the applicant has a relative that died from specified diseases) and *Employment_Info_1* (continuous variable about annual income). The remaining variables are used as independent variables except for applicant's *ID*. The columns (attributes/variables) include information about the applicants such as their personal information (age, height, weight, etc.), employment information, insurance history, family history, and medical records. A sample of ten rows and ten columns of the data is shown in Table 1.

Ins_Age	BMI	Product_Info_1	Employment_Info_1	InsuredInfo_1	Insurance_History_1	Family_Hist_1	Medical_History_1	Medical_Keyword_1	Response
0.64	0.32	1	0.03	1	1	2	4	0	8
0.06	0.27	1	0.00	1	2	2	5	0	4
0.03	0.43	1	0.03	1	2	3	10	0	8

Table 1. An example of records in the dataset wherein each row is an applicant.

Brief data preparation details

The variables in the dataset are categorical, discrete, and continuous. From the total of 128 variables, 13 have missing values[1] (Appendix A). None of the columns with missing values were dropped because after a careful analysis, the researchers concluded that these variables (some of which are related to the outcome variables) might be very important for creating predictive models. Therefore, the missing values were filled with zeros or central tendency values as the context demanded and as guided by the correlation results. Also, this way of imputation is chosen for ease of computation and interpretability. For those with excessive missing values (at least 75% is missing) which is the case for the discrete variables: *Medical_History_10, _15, _24,* and *_32*, the researchers turned them into categorical variables wherein one of the categories pertains to the value being missing.

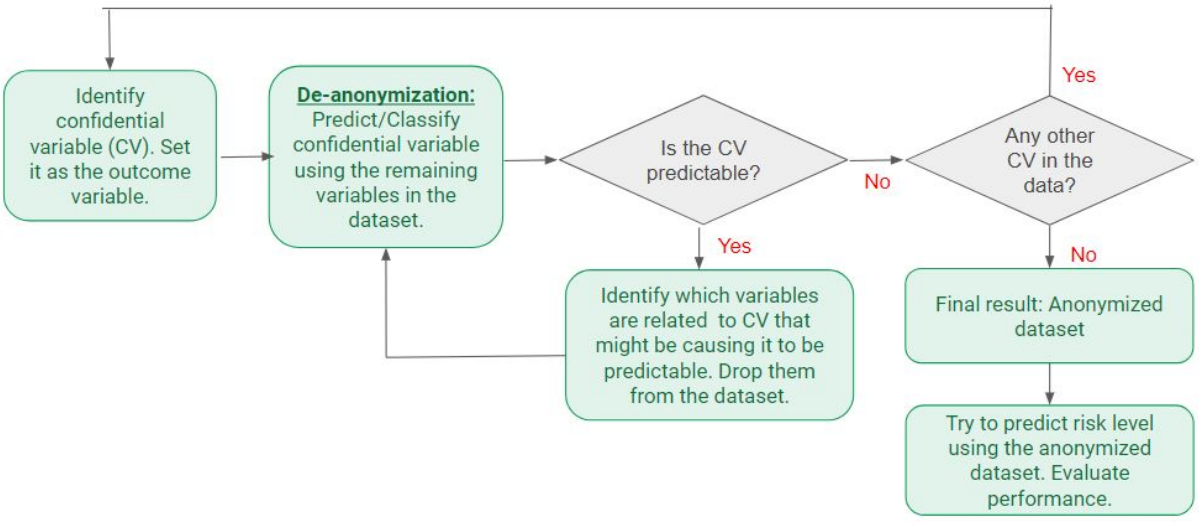


Figure 1. The flowchart above shows the methodology for the whole process of de-anonymization.

De-anonymizing Family_History_1. The death-related information and categorical variable *Family_Hist_1* (with levels 1, 2, and 3) is set as the outcome variable. Following the flowchart in Figure 1, *Family_History_1* is the chosen confidential variable (CV). Given the huge number of independent variables (over 900 in all), *Principal Components Analysis* (PCA) and *Random Forest Feature Importances* methods were used to reduce the dimension of the data. It was concluded that the results from Random Forest is more predictive and also more interpretable. Using a 5-fold cross-validation for hyper-parameter selection, it was found out that focusing on the 20 important variables is enough for building models for *Family_Hist_1*. Using these 20 variables, different multi-class classification models were tried such as Logistic regression, Decision Trees, K-Nearest Neighbors, and

[1] Missing variable labels

Ensembling methods. To evaluate model performance, the averaged version of *Accuracy*, *Precision, Recall,* and *F1-score* for multi-class classification were used. The higher the values of these metrics, the better is the performance of the model. Macro-averaging weighs all classes equally while micro-averaging weighs each prediction/datapoint equally. The best model is a Decision Tree with an accuracy of 78% and has the best trade-off for precision and recall.

Best trained (Hyperparameter tuned) models using 20 variables	Test Accuracy	Precision		Recall		F1-score	
		Micro	Macro	Micro	Macro	Micro	Macro
Logistic Regression	0.70	0.70	0.62	0.70	0.53	0.70	0.51
Ensemble (Bagging)	0.78	0.78	0.67	0.78	0.75	0.78	0.69
Decision Tree	0.78	0.78	0.65	0.78	0.80	0.78	0.68

Table 2. The table shows the three best predictive models for Family_Hist_1 using the 20 variables selected from Random Forest Feature Importances method.

Given that *Family_Hist_1* can be predicted with decent accuracy, the next step is to identify which variables highly contribute to the prediction. Based from the result of the the Decision Tree, the important features to Family_Hist_1 are shown in Figure 2.

```
1) Family_Hist_3                0.537035
2) Family_Hist_5                0.253123
3) Family_Hist_4                0.113671
4) InsuredInfo_1_1              0.071971
5) Family_Hist_2                0.014592
6) Employment_Info_1            0.005705
7) BMI                          0.002499
8) Insurance_History_5_mean     0.001004
9) Wt                           0.000400
10) Employment_Info_6           0.000000
```

Figure 2. The figure above shows the GINI criterion of the first 10 important variables to Family_Hist_1.

The variables (starting from *Family_Hist_3*) are dropped one-by-one and the metrics are evaluated to find which variables or combination of variables will make a huge drop in the performance of the model. The result of this somewhat sensitivity analysis can be checked in Appendix C. It was found out that removing *Family_Hist_3* and *Family_Hist_5* made the highest drop in terms of the metrics used, and thus, these are the most important variables in predicting *Family_Hist_1*. This makes sense because if the researchers are right in cross-referencing these variables with the application form, *Family_Hist_3* and *Family_Hist_5* are asking for the Father's and Mother's age of death. For the rest of the next analysis, the variables *Family_Hist_1, Family_Hist_3, and Family_Hist_5* are removed from the data.

De-anonymizing Employment_Info_1. Given the previous result, *Employment_Info_1* is the next CV to de-anonymize. This is a continuous variable pertaining to a normalized annual income. For convenience (and because of time constraint), the researches only tried a Decision Tree Regressor model to predict *Employment_Info_1.* Mean-squared error (MSE) is used to evaluate the performance of the model, which means the lower, the better the model is in predicting income. The best tuned model has 20 variables and a maximum depth of 5. Then doing the same sensitivity analysis as in *Family_Hist_1*, the important features (given by the best model) in predicting *Employment_Info_1* are dropped one-by-one. It was

found out that excluding *Employment_Info_6, Insurance_History_5, Product_Info_4*, and *Employment_Info_2* resulted to the highest drop in predicting *Employment_Info_1* in terms of MSE. Refer to Appendix D for the de-anonymization results of *Employment_Info_1*.

Evaluating Risk Level Prediction. The anonymized dataset is the data excluding the variables: *Family_Hist_1, Family_Hist_3, Family_Hist_5, Employment_Info_1, Employment_Info_6, Insurance_History_5, Product_Info_4*, and *Employment_Info_2.* The researchers checked if the anonymized data is still able to predict the risk level of applicants. This is important for insurance companies because they would want to have data that can be used in assessing the risk-level of applicants. The researchers found out that predicting the multi-level (8 levels) categorical variable *risk-level* of applicants is too difficult, a Random Forest model is no better than random guessing. To solve this problem, and for ease of evaluation, the risk-level variable is changed to a binary variable with levels 1 to 4 as one category and 5 to 8 as another category. This choice of differentiating the levels is based on the percentage distribution of risk level of applicants in the dataset (refer to Appendix E). Using the original dataset, a hyperparameter-tuned Random Forest model can predict the binary risk-level with 80.8% accuracy. Using the anonymized dataset, a hyperparameter-tuned Random Forest model can predict the same outcome with 80.0% accuracy. The drop in accuracy (-0.99% difference) is not huge and is not alarming for risk-level prediction. Appendix F contains the table for the binary risk-level prediction accuracy.

Conclusions

The main conclusions of the research are the following:

- Dropping the identified sensitive variables (and the important variables related to them) is possible and did not *significantly* affect the risk level prediction.
- The performance metrics (setting a threshold for de-anonymization) are critical and should be discussed with NAIC.
- The assumptions about the variables, especially the specific variable labels, should be checked with Prudential.
- Also, considering that the researchers used their own judgement to define which variables are sensitive information and which are not, a discussion with NAIC and Prudential will surely make a better guided decision and selection process.

Recomendations

As part of the final analysis, considering technical and practical perspectives, these are some of the recommendations:

- Repeat the de-anonymization process with the remaining identified sensitive variables. The current research considered only two outcome variables from the identified list which did not make a huge effect in the accuracy level of the risk prediction.
- Evaluate the risk-level modelling everytime new variables are excluded from the dataset.
- The NAIC should make a formal research (or survey) in identifying sensitive variables in the current insurance application process.

Appendix

Appendix A

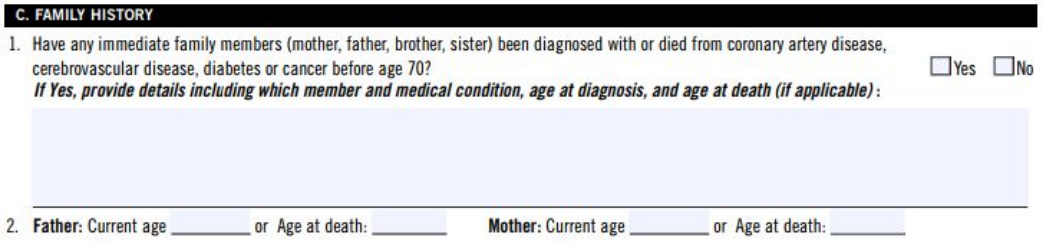
Out of 128 columns, 13 has missing values. The table below shows these 13 variables and the their percentage of missing values.

Variables	Count of Misisng Values	%
Employment_Info_1	19	0.03
Employment_Info_4	6,779	11.4
Employment_Info_6	10,854	18.3
Insurance_History_5	25,396	42.8
Family_Hist_2	28,656	48.3
Family_Hist_3	34,241	57.7
Family_Hist_4	19,184	32.3
Family_Hist_5	41,811	70.4
Medical_History_1	8,889	15.0
Medical_History_10	58,824	99.1
Medical_History_15	44,596	75.1
Medical_History_24	55,580	93.6
Medical_History_32	58,274	98.1

Appendix B

Cross-referencing the variables with Prudential's life insurance application form.

1. Family History



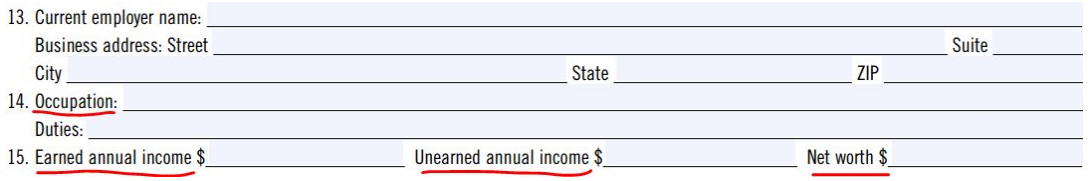
C. FAMILY HISTORY

1. Have any immediate family members (mother, father, brother, sister) been diagnosed with or died from coronary artery disease, cerebrovascular disease, diabetes or cancer before age 70? ☐Yes ☐No
   ***If Yes, provide details including which member and medical condition, age at diagnosis, and age at death (if applicable)*:**

2. **Father:** Current age _____ or Age at death: _____ **Mother:** Current age _____ or Age at death: _____

The sensitive variables were identified from the application form, where question 1 asks details about the death of the relatives of the applicant. For *Family History*-related variables, through exploration results and correlation analysis, the researchers assume that *Family_Hist_1* refers to question 1, while *Family_Hist_2* is the Father's current age, *Family_Hist_3* is the Father's age of death, *Family_Hist_4* is the Mother's current age, and *Family_Hist_5* is the Mother's age of death.

2. Employment History


13. Current employer name: _____
Business address: Street _____ Suite _____
City _____ State _____ ZIP _____
14. Occupation: _____
Duties: _____
15. Earned annual income $_____ Unearned annual income $_____ Net worth $_____

The researchers hypothesized that *Employment_Info_1*, *Employment_Info_4*, and *Employment_Info_6* may pertain to earned annual income, unearned annual income, and net worth which are all continuous variables.

Appendix C

The table shows the result when the important features in predicting Family_Hist_1 are excluded in the model one-by-one.

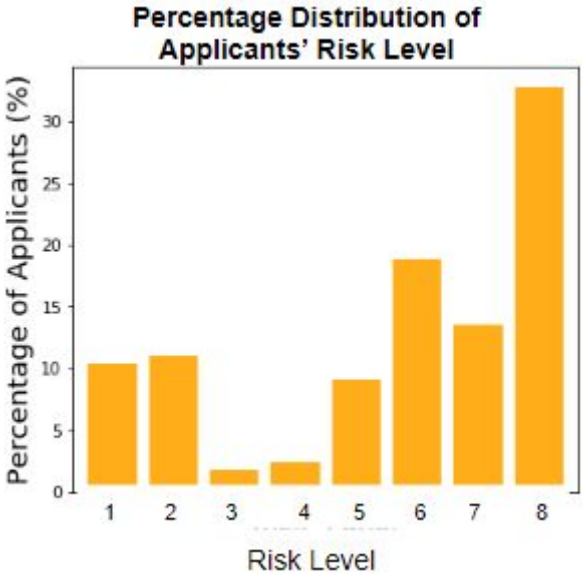
Excluded Variables	Accuracy	Precision		Recall		F1-score	
		Micro	Macro	Micro	Macro	Micro	Macro
Nothing	0.776	0.78	0.65	0.78	0.8	0.78	0.68
Family_Hist_3	0.736	0.74	0.62	0.74	0.68	0.74	0.57
Family_Hist_5	0.758	0.76	0.48	0.76	0.44	0.76	0.45
Family_Hist_4	0.776	0.78	0.49	0.78	0.47	0.78	0.48
InsuredInfo_1_1	0.775	0.78	0.65	0.78	0.8	0.78	0.68
Family_Hist_3, Family_Hist_5	0.705	0.7	0.42	0.7	0.36	0.7	0.33
Family_Hist_3, Family_Hist_5, Family_Hist_4	0.692	0.69	0.38	0.69	0.34	0.69	0.31
Family_Hist_3, Family_Hist_5, Family_Hist_4, InsuredInfo_1_1	0.692	0.69	0.38	0.69	0.34	0.69	0.31

Appendix D

The results below show the de-anonymization for *Employment_Info_1*.

Excluded Variables	MSE	
	Train	Test
Nothing (Best DT Regressor)	0.000	0.000
Employment_Info_6, Insurance_History_5, Product_Info_4, Employment_Info_2_1	0.006	0.006

Appendix E



Appendix F

The table shows the model performance for binary risk-level prediction with and without the sensitive variables (and the variables related to them).

Outcome variable	Variables used	Test Accuracy
Binary Risk Level (0: Level 1 to 4; 1: Level 5 to 8)	(a) RF 40 variables and depth of 5	80.8
	(b) = (a) + Exclude Family_Hist_1,_3, and _5	80.8
	(b) + Exclude Employment_Info_6, Insurance_History_5, Product_Info_4, Employment_Info_2_1	80.0