De-anonymization of Insurance Applicants' Sensitive Information

Team 3: Jay Lee, Maxim Castaneda, Rosalie Dolor

Goal

To enhance insurance companies best practices by ensuring the clients' privacy rights in the information gathering process.

Dangers

Margin of error in risk predictions might increase instead of decrease.

Success

Increase in the number of applicants.

Benefits to Stakeholders

NAIC	US insurance Companies	Insurance Applicants
The results of the research will provide guidance for new insurance policies.	A higher probability of attracting potential clients. To gain new knowledge from the data mining outputs about insurance dataset.	Reduced privacy invasion.

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: _____

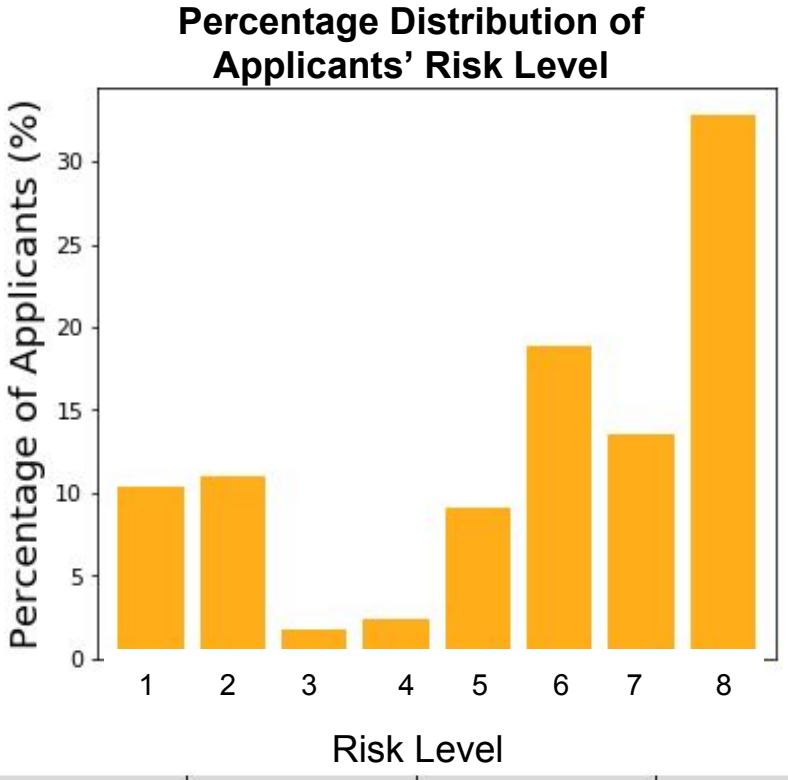
14. Occupation: _____

Duties: _____

15. Earned annual income $_____ Unearned annual income $_____ Net worth $_____

Data Description

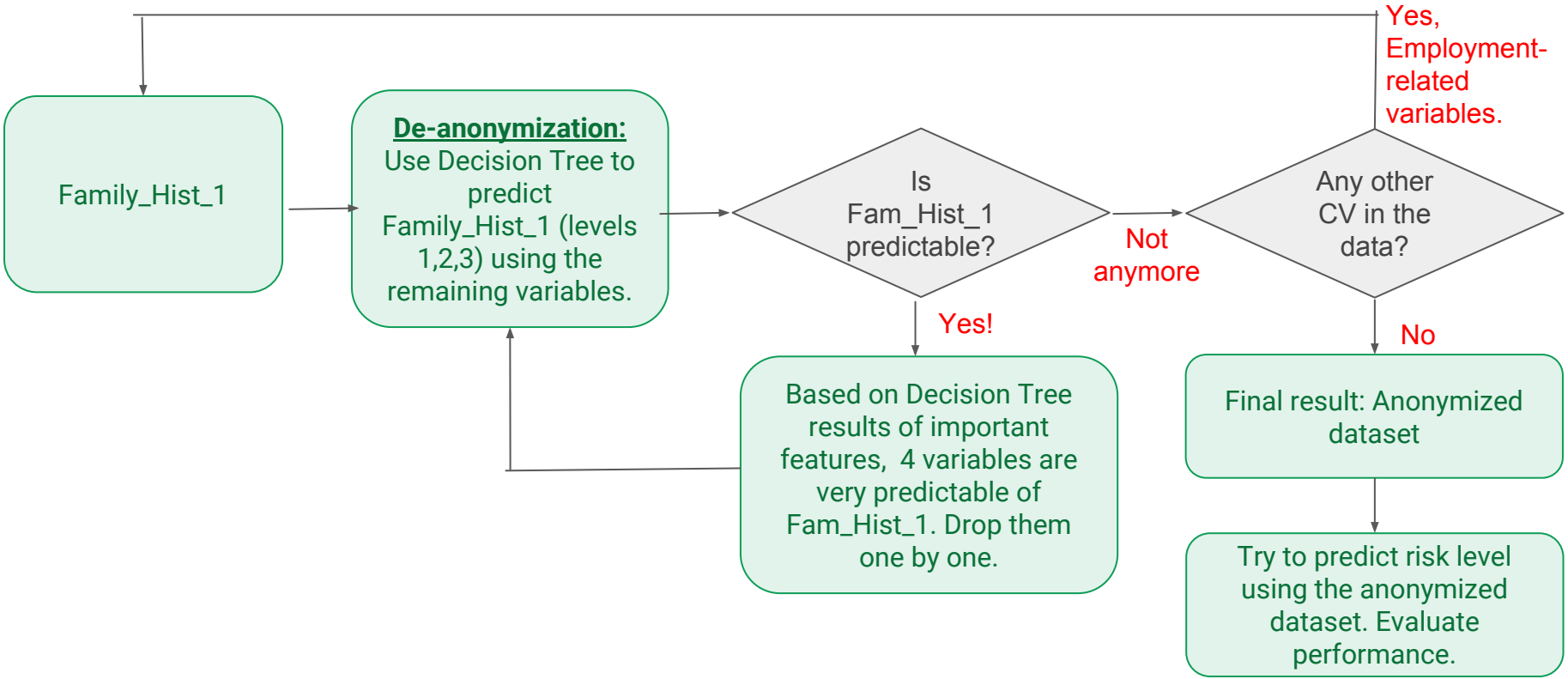
- Data source: Kaggle competition
- Size: 59,381 rows and 128 columns (with dummy variables: 900+ columns)
- Each row is an insurance applicant.
- Pre-processing and Exploration:
  - Fill in missing values
  - Correlations
  - PCA
- Partitioning:
  - Train & Test: 70%-30%
  - 5-fold Cross-validation (parameter-tuning)



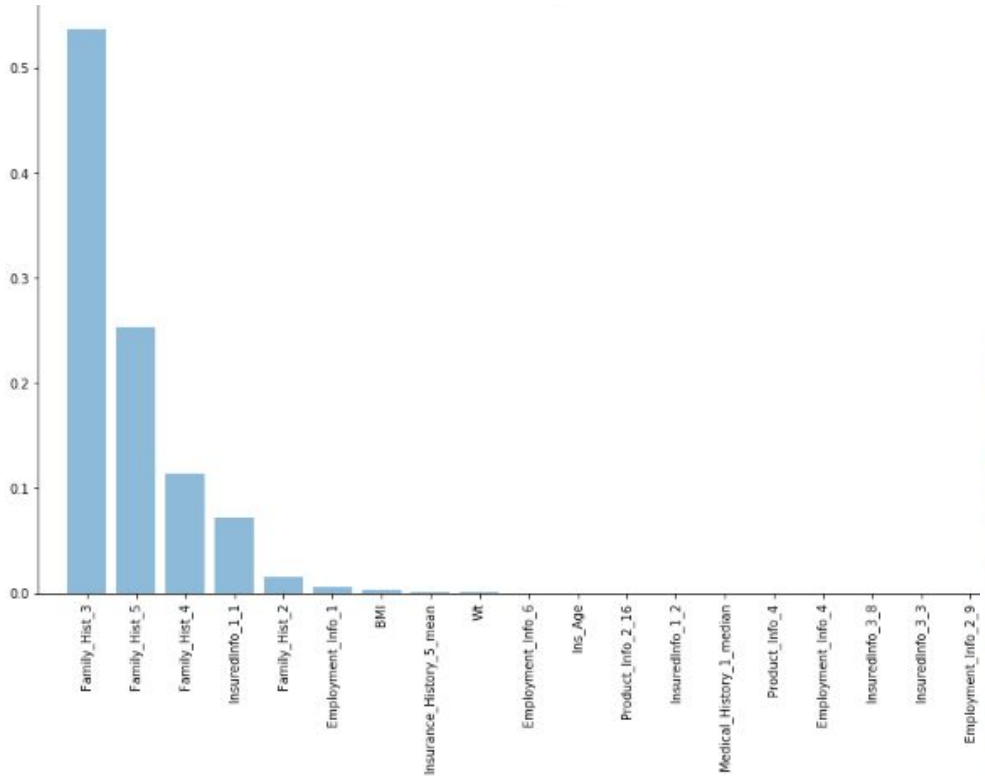
Ins_Age	BMI	Product_Info_1	Employment_Info_1	InsuredInfo_1	Insurance_History_1	Family_Hist_1	Medical_History_1	Medical_Keyword_1	Response
0.64	0.32	1	0.03	1	1	2	4	0	8
0.06	0.27	1	0.00	1	2	2	5		4
0.03	0.43	1	0.03	1	2	3	10		8

Risk Level

Methodology (Process Flow)



Feature Importances




```
 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
11) Ins_Age                        0.000000
12) Product_Info_2_16              0.000000
13) InsuredInfo_1_2                0.000000
14) Medical_History_1_median       0.000000
15) Product_Info_4                 0.000000
16) Employment_Info_4              0.000000
17) InsuredInfo_3_8                0.000000
18) InsuredInfo_3_3                0.000000
19) Employment_Info_2_9            0.000000
20) Ht                             0.000000
```

De-anonymizing Family_Hist_1

Excluded Variables	Accuracy	Precision		Recall		F1-score	
		Micro	Macro	Micro	Macro	Micro	Macro
Nothing	0.78	0.78	0.65	0.78	0.80	0.78	0.68
Family_Hist_3	0.74	0.74	0.62	0.74	0.68	0.74	0.57
Family_Hist_5	0.76	0.76	0.48	0.76	0.44	0.76	0.45
Family_Hist_4	0.78	0.78	0.49	0.78	0.47	0.78	0.48
InsuredInfo_1_1	0.78	0.78	0.65	0.78	0.80	0.78	0.68
Family_Hist_3, Family_Hist_5	0.71	0.70	0.42	0.70	0.36	0.70	0.33
Family_Hist_3, Family_Hist_5, Family_Hist_4	0.69	0.69	0.38	0.69	0.34	0.69	0.31
Family_Hist_3, Family_Hist_5, Family_Hist_4, InsuredInfo_1_1	0.69	0.69	0.38	0.69	0.34	0.69	0.31

De-anonymizing 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

Evaluating the Risk Level Prediction

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

Implementation/Production Considerations

NOTES

- Assumptions about the variables should be checked with Prudential.
- Based on the results, dropping the identified sensitive variables (and the important variables related to them) is possible and it did not *significantly* affect the risk level prediction.
- Performance metrics (setting a threshold for de-anonymization) is critical and should be discussed with NAIC.

RECOMMENDATIONS

- Repeat the algorithm with the remaining identified sensitive values.
- Re-evaluate risk level modelling.