The Report Committee for Xinqi You

Certifies that this is the approved version of the following report:

A Hybrid Reduced Approach to Handle Missing Values in Type 2 Diabetes Prediction

APPROVED BY

SUPERVISING COMMITTEE:

_____

Maytal Saar-Tsechansky, Supervisor

_____

Kishore Gawande

A Hybrid Reduced Approach to Handle Missing Values in Type 2 Diabetes Prediction

by

Xinqi You, B.S.

REPORT

Presented to the Faculty of the Graduate School of
The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

MASTER OF SCIENCE IN STATISTICS

The University of Texas at Austin
May 2016

Abstract

A Hybrid Reduced Approach to Handle Missing Values in Type 2 Diabetes Prediction

By

Xinqi You, M.S. Stat.
The University of Texas at Austin, 2016
Supervisor: Maytal Saar-Tsechansky

Diabetes gains more attention among medical institutions and health care organizations as the increasing trend of diabetes around the world. In the United States, 29.1 million people or 9.3% of U.S. population are diagnosed with diabetes. About 86 million people are categorized as pre-diabetes and 15-30% of them will develop diabetes within 5 years. To tackle this challenge, National Diabetes Prevention Program (DPP) was introduced in 2002 and it reduces risk of diabetes by 58% through lifestyle change program. In order to help select a better group of pre-diabetes for intervention and maximize the cost-effectiveness of the program, we propose a Hybrid Reduced approach to handle missing values when predicting type 2 diabetes. This approach deals with 4 challenges in electronic medical records: missing values, missing not at random, class imbalance and predicting at a longer window (2-year). We select three ensemble predictive models: AdaBoost.M1, Gradient Boosting and Extremely Randomized Trees and apply this approach across 7 years to assess its robustness. The Hybrid Reduced approach includes two sub-approaches: Hybrid Reduced Organic and Hybrid Reduced Imputed. Throughout the experiments, Hybrid Reduced Imputed is the best performer and achieves a 5-7% improvement in precision. By simply using this approach, we could save $278 million for healthcare and improve people's health condition.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1. Predictive Analysis in Diabetes

As more healthcare data are being collected and made public, predictive analytics in healthcare has risen to be a hot topic in recent years. One specific field of public interest is in diabetes. In United States, it is estimated that there are 29.1 million or 9.3% of the population have diabetes by 2014, 72.2% of them are being diagnosed. During 2009-2012, 37% of U.S. adults aged 20 years or older had pre-diabetes based on their fasting glucose or A1C levels, which is approximately 86 million people. Among the pre-diabetes, 90% of them do not know they have pre-diabetes and 15-30% of them will develop type 2 diabetes within 5 years. More importantly, the risk of death for adults with diabetes is 50% higher than the adults without diabetes.

In 2012, the estimated diabetes costs in the United States are $245 billion including $176 billion direct medical costs and $69 billion indirect costs by American Diabetes Association. Care for people with diagnosed diabetes takes up about 20% of U.S. total health care. On average, people with diagnosed diabetes spend about $13,700 per year on medical treatment, of which around $7900 directly attributes to diabetes. The medical expenditure of diabetes is 2.3 times higher than those of non-diabetic people.

Given such large diabetic population and medical expenditure, it is critical for both medical institutions and health insurance companies to help prevent diabetes before it takes place. With this purpose in mind, Centers for Disease Control and Prevention started the National Diabetes Prevention Program (DPP) in 2002. DPP and its partner organizations work to reduce the growing problem of pre-diabetes and type 2 diabetes. The key component of DPP is the Lifestyle Change Program. It helps pre-diabetes people prevent or delay type 2 diabetes. The DPP Lifestyle Change Program is highly successful with a 58% reduction in the risk of diabetes through a study of 3234 non-diabetic people with elevated fasting glucose level

(DPP Research Group, 2006).

A widely used approach to identify pre-diabetes patients is to use the Diabetes Risk Score (Lindström, et al. 2003), which includes age, BMI, waist circumference, history of antihypertensive drug treatment and high blood glucose, physical activity, and daily consumption of fruits or vegetables. The Score is derived from a logistic regression based on previous questionnaire. There are other risk scores based on similar approaches are also very popular. Another direction in identifying type 2 diabetes is to examine its genome-side association. Some variants of genes are associated with type 2 diabetes (Scott, Laura J., et al. 2007).

With more electronic medical records available, we want to improve diabetes prediction and help select more high-risk pre-diabetes patients into the prevention group so as to reduce diabetes through machine learning.

1.2. Challenges in Predicting with Electronic Medical Data

Recently there are many studies with electronic medical data in machine learning and data mining. Specifically in type 2 diabetes prediction, relative research in predicting diabetes or blood glucose levels are focused on feature selection (Huang, Yue, et al. 2007), applications of machine learning classification algorithms such as Supporter Vector Machine (Yu, Wei, et al. 2010), or comparing different machine learning algorithms (Mani, Subramani, et al. 2012).

Though with massive amount of electronic medical data available, there are some challenges while working with them. Also, the prediction for type 2 diabetes is also at a much shorter window (3 months to 1 year). The Diabetes Prevention Program usually runs for several years and it's a long-term effort in preventing diabetes.

1.2.1. Missing Values and Missing Pattern

The first problem is the existence of a lot of missing values. Missing values are very common in health research. A comprehensive review on handling missing data

in diabetes risk prediction models shows that only 37.5% studies mentioned reporting on percentage of missing values in their datasets (Masconi, Katya L., et al. 2015). 45.8% chose to delete records with missing values and 10.4% use imputation methods. The rest either use complete datasets or do not report their treatment for missing values. Data size is also relatively small among the existing research, at about 6900 records on average. This lack of information might due to poor data management or the studies use more historical data, when electronic medical records were not as popular as of now.

According to the review, none of the selected articles in diabetes risk prediction models discussed patterns of missing data or provided reasons for the missing data. There are three types of missing data, missing completely at random (MCAR), missing at random (MAR) and missing not at random(MNAR). MCAR means no systematic difference between the missing and observed values. The reason for missing is completely random, such as the machine breaks or the nurse takes a leave. In contrast, MAR assumes the systematic difference between the missing values and the observed values could be explained by the observed data (Sterne, Jonathan AC, et al. 2009). The missing blood sugar measurement might because physicians decide not to conduct the test since these patients are young and not overweight. MNAR is that even after the observed data are taken into account, systematic difference between missing and observed values still exist. For example, people with severe headache or depression are more likely to miss clinical appointments.

In electronic medical records, missing values usually come from certain variables with lab tests or care information. When patients do not take a lab test, it is barely due to complete random decision that they are unwilling to take the test. It could be that physicians did not require those tests (MAR) or they consider themselves healthy and unnecessary to check (MNAR). It is difficult to classify whether it is MAR or MNAR and even chances are that the data is mixed with MAR and MNAR.

In the current studies of handling missing data, existing approaches include deletion, single imputation, maximum likelihood estimation, Bayesian estimation, multiple imputation (Enders 2010) and reduced-feature models (Saar-Tsechansky & Provost 2007). There are many variations based on these methods. Notice that for most of the above methods, they assume either MCAR or MAR. Although MAR-based methods are the current state of art (Schafer & Graham, 2002) , MNAR still raises more attention especially with longitudinal and clinical trial data (Pauler et al. 2003). When dealing with MNAR data, generally there are selection models and pattern mixture models.

For most imputation or estimation methods, the imputed or estimated values are based on observed values. It could introduce potential bias or incorrect values for the missing data. Therefore, we found the reduced-feature models more appealing since it does not require imputation or estimation and only build models based on the observed values.

1.2.2. Predicting at a Longer Window

The second challenge is to predict at a longer window. Previous studies usually predict in a window of 3 months to 1 year, in which the patients might already be diabetic at that time. Even for people with severe gain in BMI between age 25 and 55, it still takes 1.95 – 3.91 years before they develop type 2 diabetes (Schienkiewitz, Anja, et al. 2006). Therefore, predicting shorter than 2 years barely provides practical insights for diabetes prevention institutions. And this also partly explains the high accuracy or sensitivity in previous diabetes risk prediction. The diabetes prevention programs are a long-term effort and it aims to change lifestyle, which indeed is a very gradual process.

1.2.3. Class Imbalance

The third challenge is the class imbalance problem. A dataset is class imbalanced if the classification categories are not equally represented. Usually the

minority class is of special interest in classification (Chawla et al. 2004). Although the percentage of diabetes increases in the past few years, it is still around 8% of the total population. In this case, even if we gain 92% accuracy in prediction, we might still have 0% in recall. Also the 8% comes from the estimate of general population, for different medical institutions, the percentage also varies.

There are mainly four subareas to tackle class imbalance problem: sampling, one-class learning, feature selection and ensemble learning. Sampling methods include under-sampling majority class, over-sampling minority class or the combination of them. The idea is to balance the datasets through sampling techniques. It is usually combined with ensemble algorithms such as RusBoost (Seiffert et al. 2010) and SMOTE (Han et al. 2005) respectively. One-class learning (Kubat et al. 1997) is a recognition-based approach that provides alternative discrimination towards the class not of interest. The feature selection proposes that use different features for positive and negative classes and then explicitly combine them (Zheng et al. 2004). Ensemble learning aims to improve the performance of single classifiers by including multiple classifiers and combining them to obtain a new classifier. Ensemble learning main focuses on boosting and bagging as well as their numerous variations cater to different types of datasets (Galar et al. 2012).

2. Approaches to Treat Missing Values

2.1. Previous Work on Reduced-Feature Model in Handling Missing Values

Reduced-feature models are based on the intrinsic characteristic of the missing patterns within datasets. It only includes attributes that are known when the predicting at the test instances. Therefore, a new classification model is trained after removal of features that are not present in test set (Saar-Tsechansky & Provost 2007). This approach takes advantage of the "naturally missing" patterns and does not assume either MAR or MNAR.

2.2. Revised Reduced Approach

Our datasets have a large amount of missing values. The existence of missing values will greatly decrease the performance in predictive analysis. Besides the standard approaching of imputing all missing values (Complete Imputed), we propose two versions of Reduced approach that will take the missing not at random into consideration. Imputations are usually based on the observed values and a lot of times there is significant bias when we are imputing unknown values.

2.2.1. Complete Imputed

Complete Imputed is defined as imputing all missing values in both training data and testing data. Imputation methods vary based on the preference or requirements. In this report, we will use multiple imputation, as it is the state-of-the-art imputation approach. In modeling phase, this approach will train on all records with imputed values and test on entire test set with imputed values.

2.2.2. Reduced Organic

Previous research on reduced approach to treat missing data is to segment dataset into subsets based on its missing patterns. Subsets with the same missing pattern will be grouped together and build models on these subsets. In the modeling part, the subsets of the train and test sets with the same missing pattern will only

use the available features, which are mutual in the train and test sets. In this way there is no imputation and it shows significant advantage in taking naturally missing patterns.

Based on this approach and the characteristics of our dataset, we propose the Reduced Organic approach that we still build separate models for subsets with naturally missing patterns, but the criterion for qualifying a subset is that the pattern has at least 100 records in the training dataset (which is around 0.15% of data size). The records that fail to meet this criterion will be put into the complete model, which uses the Complete Imputed approach. The reason why we have this criterion here is to ensure there are enough observations in the training set and to reduce potential variance.

For example, if there are *k* missing patterns that meet the criteria stated above, Reduced Organic approach selects *k* subsets in test set that corresponds to the missing patterns as in train set. Records in train and test set that do not belong to the *k* missing patterns will use Complete Imputed approach and implement predictive models on the imputed datasets. Then it will build *k+1* models and output *k+1* lists of probability rankings for each model.

2.2.3. Reduced Imputed

Instead of using only subsets of missing patterns in the training sets, Reduced Imputed will perform imputation on the entire train set and use all records with subsets of features in modeling phase. Here we still need to obtain the subsets of missing patterns in the test set. With the same selection criteria as in Reduced Organic, we select subsets corresponding to their missing patterns in test sets. Records that are not in subsets are put together and impute missing values in this subset.

In the modeling phase, we build *k* models that use the entire imputed train sets with features corresponds to the *k* missing patterns. Then test on the subsets of test sets with the same missing patterns. There is no imputation on test sets. For the 'left

over' subset, we build a model on the entire imputed train sets with all features and test on the 'left over' subset of test set with all features. Therefore, we will have *k+1* models and *k+1* lists of probability rankings for each model.

2.3. Hybrid Reduced Organic and Hybrid Reduced Imputed

To compensate the fact that less frequent patterns have few records (<1% of total records), we propose a more hybrid approach for Reduced Organic and Reduced Imputed. Instead of selecting all *k* missing patterns that have more than 100 records, we limit *k* to 20, which only takes more frequent missing patterns in the train sets. This number is determined by the characteristic of our dataset, which at that point train set will have adequate data. Other users could use the threshold based on their datasets.

Meanwhile, we cross-validate the performance from reduced approaches with Complete Imputed approach and decide whether to use reduced approaches. If the reduced approaches in this particular missing patterns is better than Complete Imputed, we keep the results of reduced approaches, otherwise those records in both train and test sets corresponds to this pattern will use Complete Imputed approach later on. Then for the non-frequent-missing-pattern records and records that reduced approaches perform worse than Complete Imputed approach, we apply Complete Imputed approach.

For example, for top frequent missing patterns; if *m* patterns ($m \leq k$) have better performance in reduced approaches, we will have *m+1* models in the end for Hybrid Reduced Organic and Hybrid Reduced Imputed.

The hybrid approach gives advantage to reduced approaches for more training records in the model and incorporate standard Complete Imputed for less frequent naturally missing records. Ideally we would like to do the cross-validation for every pattern, for the reason of efficiency we only select *k* patterns. Also we saw that as the patterns become less frequent, the bias gets very larger.

3. Experiment and Results

3.1. Experimental Setup

In order to examine the performance of Reduced Organic and Reduced Imputed, whether using Hybrid or not, we will compare their performances with Complete Imputed. The scenario for evaluation is to predict diabetes on a two-year window. We have 10 consecutive years of electronic medical data, which are from the same organization with exact same features. The electronic medical data of each year contains around 65000 patient information, which include physical biometrics (blood pressure, temperature, etc.), lab test results

The latest dataset lacks the labels for diabetes of the future so we end up with 7 pairs of train and test sets. Train sets are Year1 to Year7 and test sets are Year3 to Year 9 respectively. Each dataset (Year1 to Year10) contains the binary labels for whether these patients develop diabetes at the time point two years later.

The intuition for using train and test datasets is that when we have real-world datasets at the time for prediction, we do not know whether the labels for them. The available sources for training a predictive model is from historical datasets with labels. Medical data are different from other data in a way that the window is much longer (across decades) and more drastic variations during the time period. The traditional approach of splitting datasets into train and test sets fails to take into account the time-varying effects across years.

The experiment is repeated 7 times in the time-moving window, which is a further cross-evaluation of the effectiveness on reduced approaches. This is also essential in real-world practice if medical institutions want to evaluate their predictive models.

3.1.1. Feature Selection

This dataset has 107 features in total. To reduce noise and utilize most informative features, we performed feature selection by calculating the average entropy-based gain ratio of each variable when predicting on a two-year window.

This process was conducted in R with package 'FSelector'.

We calculated the gain ratio for each dataset and obtain the average gain ratio for each feature. By calculating gain ratio of each feature, we excluded features with gain ratio smaller than . Therefore, the new datasets that will be used in experiment section have 37 features.

3.1.2. Imputation Method

In order to run different models on the datasets, we need to impute missing values for Complete Imputed and Reduced Imputed approaches. In our datasets after feature selection, 73% of features have missing values over years. As shown in Figure 4.2.1, each line represents a feature. The rate of missing declines over years, which means medical institutions are collecting more and more sufficient data.

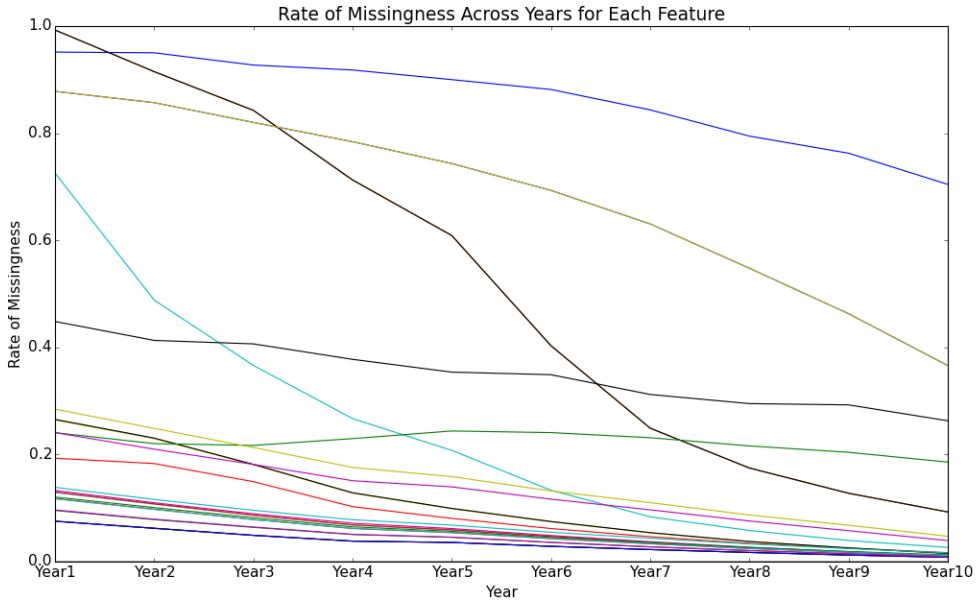

Figure 3.1 Rate of Missing Values Across Years for Each Feature

The imputation method we used is Multiple Imputation by Chained Equations (MICE). MICE has emerged as the state-of-art imputation method when dealing with

missing values. It creates multiple imputations instead of single imputation and the chain equations approach handles different data types. Note that MICE assumes Missing at Random (MAR), but in our case, the missing is not at random. However, in order to implement predictive models, MICE is the most effective imputation methods. We used Microsoft Azure Machine Learning platform to perform MICE on all datasets with 5 iterations.

3.1.3. Prediction Models

In prediction models, we select three ensemble models: AdaBoost SAMME, Gradient Boosting and Extremely Randomized Trees. We mainly select these ensemble predictive models because they perform relatively better in practice. The predictive models in the experiments are based on the Python 2.7 Scikit-Learn module (version 0.17.1).

Note that specifications of the parameters are not critical in this report and we mainly use default parameters as specified in the scikit-learn module. Therefore, we only compare results within the predictive models and not across them.

3.1.4. Measurement

The measurement used in this report is precision. The initiative of this measure is based on the goal of selecting more 'accurate' pre-diabetic patients into the intervention group. Since the labels are highly imbalanced (2-4% positive class) and with the constraints of potential budgets for intervention program, we only select top 2% of total population for intervention programs. In practice, these 2% patients are labeled as pre-diabetic patients and will be introduced to intervention groups or coaching programs that help them prevent diabetes at the early stages.

Precision is defined as by taking top 2% in the probabilities ranking from predictive models, the rate of true positive (true diabetic). The reason why we take 2% comes from the constraints of medical availability and budget.

Precision = Num. of True Positive in 2% Population / 2% Population Size

3.1.5. Hypothesis Testing

To examine if there's statistical difference among approaches and avoid potential bias in datasets, we conducted 10-time cross validation by randomly selecting 80% of train set and 80% of test set at each time. Then use paired t-test to see statistical significance among Complete Imputed, Reduced Organic and Reduced Imputed.

3.2. Evaluation of Reduced Organic and Reduced Imputed

3.2.1. Precision

First we compare the precisions of Complete Imputed, Reduced Imputed and Reduced Organic. Results in Figure 3.2-3.4 are the average precision of the 10-time cross validation. We can see that precisions of Reduced Imputed and Reduced Organic are very close to Complete Imputed in all three predictive models. They are very comparable and sometimes reduced approaches are worse than Complete Imputed.

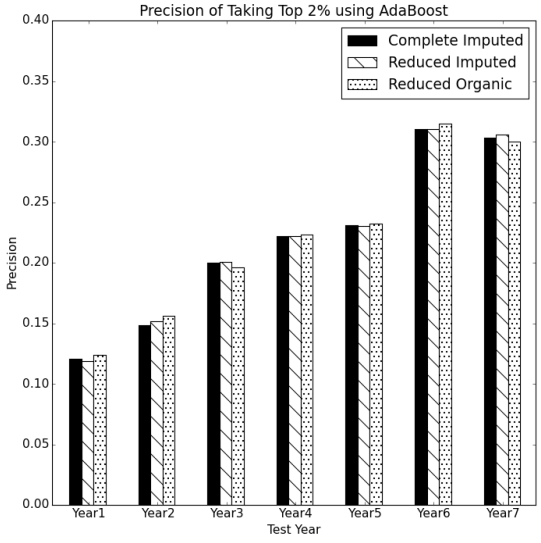

Figure 3.2 Precision Comparison of existing approaches in AdaBoost

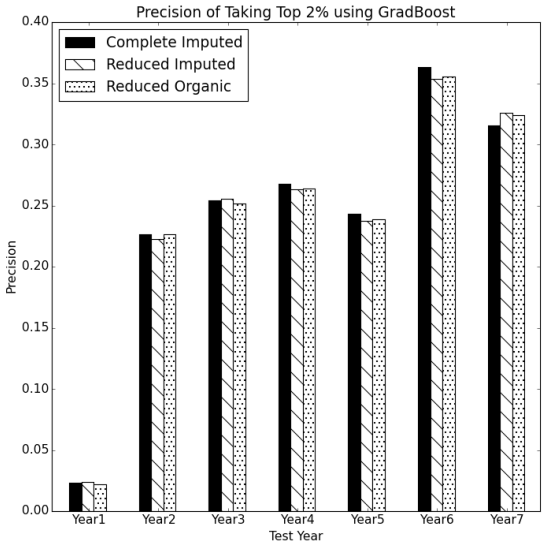

Figure 3.3 Precision Comparison of existing approaches in Gradient Boosting

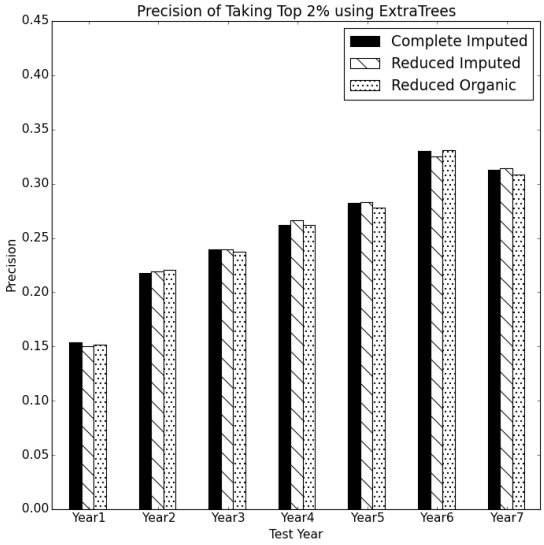

Figure 3.4 Precision Comparison of existing approaches in Extra Trees

3.2.2. Hypothesis Testing

From the paired t-test, no statistical significance was found when comparing precisions of Complete Imputed and Reduced Imputed. In comparison of Complete

Imputed and Reduced Organic, only when predicting at Year 2 the difference is statistically significant at 10% and the improvement in precision is a mere 0.78%. Therefore, three approaches are not statistically different in precision in most occasions.

Part of the reason why reduced approaches not seem to work here is that for less frequent missing patterns, few records are available in the train set. When building a predictive model on a small data size, it will introduce larger variance and bias. Even for the pattern that has hundreds of observation, it sill only accounts for few percentage of the entire dataset. Also, some patterns experience more class-imbalance problem (only <0.01% positive class in train set). Still using the threshold of 2% regardless of this problem will further hurt the performance. Based on this result, we further examine the effectiveness in Hybrid approaches.

	Precision		
Test Year	AdaBoost	Gradient Boosting	Extra Trees
Year1	0.28%	-0.14%	-0.24%
Year2	0.78%*	0.01%	0.29%
Year3	-0.38%	-0.21%	-0.20%
Year4	0.14%	-0.37%	0.00%
Year5	0.14%	-0.42%	-0.48%
Year6	0.50%	-0.77%	0.10%
Year7	-0.28%	0.80%	-0.41%

Table 3.1 Difference of Reduced Organic and Complete Imputed in CV Results

	Precision		
Test Year	AdaBoost	Gradient Boosting	Extra Trees
Year1	-0.25%	0.04%	-0.37%
Year2	0.33%	-0.38%	0.15%
Year3	0.09%	0.14%	0.00%
Year4	0.05%	-0.44%	0.42%
Year5	-0.04%	-0.53%	0.04%
Year6	0.00%	-0.94%	-0.50%
Year7	0.31%	1.01%	0.17%

Table 3.2 Difference of Reduced Imputed and Complete Imputed in CV Results
* Significant at 10% level, ** significant at 5% level, *** significant at 1% level

3.3. Evaluation of Hybrid Reduced Organic and Hybrid Reduced Imputed

3.3.1. Precision

Experiments on the Hybrid approaches are performed in a similar way as above and using the same measurements. Figures 3.5-3.47 are the average precisions from three predictive models with three approaches. We can see that Hybrid Reduced Organic and Hybrid Reduced Imputed constantly outperform Complete Imputed. Also, the Hybrid Reduced Imputed almost always performs the best, only with one exception in Extra Trees model when predicting at Year 3.

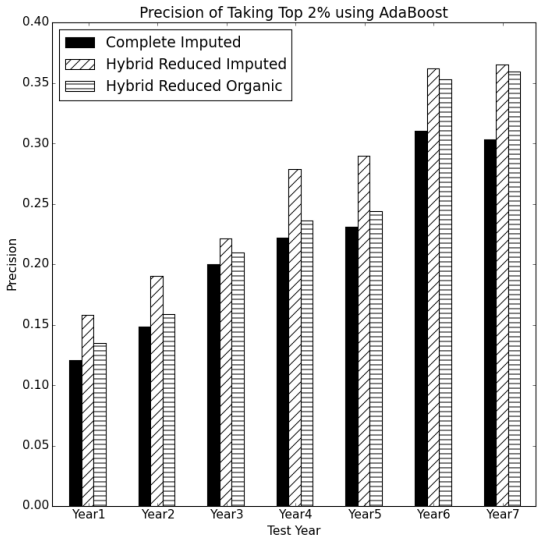

Figure 3.5 Precision Comparison of Hybrid approaches in AdaBoost

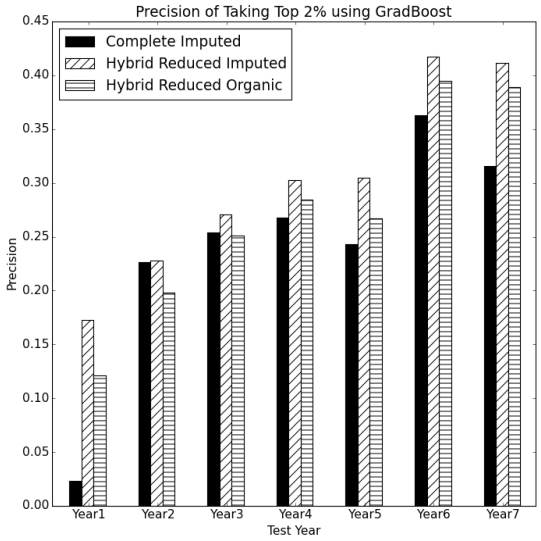

Figure 3.6 Precision Comparison of Hybrid approaches in Gradient Boosting

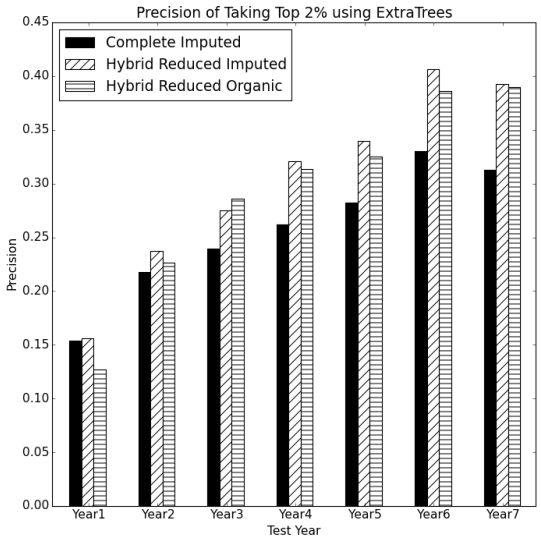

Figure 3.7 Precision Comparison of Hybrid approaches in Extra Trees

3.3.2. Hypothesis Testing

To examine the statistical significant difference among approaches, we also conducted paired t-test. Table 3.3-3.4 show the differences between reduced approach and complete approach and their statistical significance.

First we will compare Hybrid Reduced Organic and Complete Imputed. From Table 3.3, we can also see that Hybrid Reduced Imputed is almost always better than Complete Imputed with statistical significance. It shares similar characteristics as the comparison between Hybrid Reduced Imputed and Complete Imputed but in a smaller scale.

Then we compare Hybrid Reduced Imputed and Complete Imputed. In Table 3.4, Hybrid Reduced Imputed is better than Complete Imputed at 1% significance level almost at all times, with only two exceptions in Extra Trees at Year1 and Gradient Boosting at Year2. The improvement in performance also gets larger when predicting at more recent years. This is partly because as time gets more recent, there are fewer missing values and less imputation in the train datasets.

	Precision		
Test Year	AdaBoost	Gradient Boosting	Extra Trees
Year1	1.39%**	9.80%***	-2.69%***
Year2	1.01%*	-2.81%***	0.91%
Year3	0.96%	-0.31%	4.63%***
Year4	1.42%*	1.70%***	5.15%***
Year5	1.33%***	2.40%***	4.29%***
Year6	4.32%***	3.17%***	5.60%***
Year7	5.65%***	7.32%***	7.69%***

Table 3.3 Difference of Hybrid Reduced Organic and Complete Organic in CV Results

	Precision		
Test Year	AdaBoost	Gradient Boosting	Extra Trees
Year1	3.71%***	14.95%***	0.19%
Year2	4.18%***	0.11%	1.96%***
Year3	2.14%***	1.64%***	3.60%***
Year4	5.69%***	3.52%***	5.82%***
Year5	5.89%***	6.18%***	5.69%***
Year6	5.18%***	5.39%***	7.66%***
Year7	6.21%***	9.60%***	8.02%***

Table 3.4 Difference of Hybrid Reduced Imputed and Complete Imputed in CV Results

* Significant at 10% level, ** significant at 5% level, *** significant at 1% level

There are four factors that contribute to the improvement in Hybrid Reduced Organic and Hybrid Reduced Imputed.

1) Reduced approaches that emphasize less on imputation. As in the Complete Imputed, features that are potentially good predictors with missing values are imputed with possible incorrect values.

2) More records are available in the train sets in Hybrid reduced approaches since we only take more frequent missing patterns. Smaller datasets might fail to include important information and hence worse results. This advantage is very significant in Hybrid Reduced Imputed.
3) Cross-validation with Complete Imputed and only select missing patterns that perform well in reduced approaches. There are patterns with good predictors as well as bad predictors. The cross-validation screens out the bad patterns with bad predictors and puts them into the general pool.
4) No imputation in hybrid reduced test sets except for the general pool with Complete Imputed approach potentially incorporates the time-varying information. Imputing the missing values will generally reduce the performance.

Yet we do not know which factor contributes the most in the improvement, but four of them collectively lead to relatively better results than Complete Imputed.

Another thing we need to notice is that as time goes on, general precision becomes better and more significant also in Complete Imputed approach. The improvement in Hybrid Reduced Imputed is around 5 - 7% in precision starting from Year 4. The increase benefits from the more complete datasets as medical institutions are making more efforts in collecting data in electronic medical records.

4. Conclusion

4.1. Conclusion

Dealing with missing is never an easy task. As people in both academia and industry begin to analyze real-world data that have been collected during the past decades, the requests for efficiently handling missing data are increasing. There is a considerable number of research or study into this problem particularly. The existing solutions for dealing missing not at random (MNAR) is still relatively scarce as those in missing at random (MAR).

The Hybrid Reduced approaches provide an efficient solution when the missing values are MNAR. In consideration of train data size, Hybrid Reduced Imputed is the most robust approach. If the train data size of missing patterns is sufficient, we also recommend Hybrid Reduced Organic.

Generally Hybrid Reduced Organic requires much less runtime in building predictive models as the train data sets are smaller than those of Hybrid Reduced Imputed. As fewer missing data in more recent year, Hybrid Reduced Organic and Hybrid Reduced Imputed are very comparable. Practitioners could choose these two models that best fits their goals. The improvement in precision is also robust across different ensemble predictive models and years.

4.2. Practical Implications

Selecting patients for intervention group is critical in assessing the cost-effectiveness of the Diabetes Prevention Program (DPP). According to DPP, the cost for direct medical costs of Lifestyle Change Program is about $2,322 per capita per year (Herman, William H., et al. 2013), compared with $7,900 for direct diabetes medical cost per year.

By only using the Hybrid Reduced Imputed approach when identifying pre-diabetic patient, starting from Year 4 the improvement in precision is often more than 5%. If we assume the 5% precision improvement, 2% pre-diabetic population

into prevention group and 58% diabetes risk reduction, we would save $162 per person in the prevention group, which is about $278 million if the pre-diabetic population is 86 million estimated by DPP.

There are many other intervention groups which offer prevention program with lower prices or group prevention such as YMCA. Preventing diabetes is both beneficial for patients themselves and health care institutions. Big as this challenge is, we hope further research into the predictive analysis or more advanced applied methods in real-world.

Bibliography

[1] Romanski, P. "FSelector: Selecting attributes." Vienna: R Foundation for Statistical Computing (2009).

[2] Microsoft. Microsoft Azure Machine Learning Studio. https://studio.azureml.net/

[3] Azur, Melissa J., et al. "Multiple imputation by chained equations: what is it and how does it work?." International journal of methods in psychiatric research 20.1 (2011): 40-49.

[4] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." The Journal of Machine Learning Research 12 (2011): 2825-2830.

[5] Zhu, Ji, et al. "Multi-class adaboost." Statistics and its Interface 2.3 (2009): 349-360.

[6] Cheng Li. "A Gentle Introduction to Gradient Boosting" (PDF). Northeastern University. Retrieved 14 April 2016.

[7] Hastie, T.; Tibshirani, R.; Friedman, J. H.. "10. Boosting and Additive Trees". The Elements of Statistical Learning (2nd ed.). New York: Springer. pp. 337–384. ISBN 0-387-84857-6 (2009).

[8] Geurts, Pierre, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees." Machine learning 63.1 (2006): 3-42.

[9] Centers for Disease Control and Prevention. "National diabetes statistics report: estimates of diabetes and its burden in the United States, 2014." Atlanta, GA: US Department of Health and Human Services (2014).

[10] American Diabetes Association. “The Cost of Diabetes” (2015).

[11] Diabetes Prevention Program (DPP) Research Group. "The Diabetes Prevention Program (DPP) description of lifestyle intervention." Diabetes care 25.12 (2002): 2165-2171.

[12] Lindström, Jaana, and Jaakko Tuomilehto. "The Diabetes Risk Score A practical tool to predict type 2 diabetes risk." Diabetes care 26.3 (2003): 725-731.

[13] Diabetes Prevention Program Research Group. "Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin." The New England journal of medicine 346.6 (2002): 393.

[14] Scott, Laura J., et al. "A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants." science 316.5829 (2007): 1341-1345.

[15] Huang, Yue, et al. "Feature selection and classification model construction on type 2 diabetic patients’ data." Artificial intelligence in medicine 41.3 (2007): 251-262.

[16] Mani, Subramani, et al. "Type 2 diabetes risk forecasting from EMR data using machine learning." AMIA Annual Symposium Proceedings. Vol. 2012. American Medical Informatics Association, 2012.

[17] Yu, Wei, et al. "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes." BMC Medical Informatics and Decision Making 10.1

(2010): 16.
[18] Mani, Subramani, et al. "Type 2 diabetes risk forecasting from EMR data using machine learning." AMIA Annual Symposium Proceedings. Vol. 2012. American Medical Informatics Association, 2012.
[19] Masconi, Katya L., et al. "Reporting and handling of missing data in predictive research for prevalent undiagnosed type 2 diabetes mellitus: a systematic review." EPMA Journal 6.1 (2015): 7.
[20] Sterne, Jonathan AC, et al. "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls." Bmj 338 (2009): b2393.
[21] Saar-Tsechansky, Maytal, and Foster Provost. "Handling missing values when applying classification models." (2007).
[22] Schafer, Joseph L., and John W. Graham. "Missing data: our view of the state of the art." Psychological methods 7.2 (2002): 147.
[23] Enders, Craig K. Applied missing data analysis. Guilford Press, 2010.
[24] Pauler, Donna K., Sheryl McCoy, and Carol Moinpour. "Pattern mixture models for longitudinal quality of life studies in advanced stage disease." Statistics in medicine 22.5 (2003): 795-809.
[25] Schienkiewitz, Anja, et al. "Body mass index history and risk of type 2 diabetes: results from the European Prospective Investigation into Cancer and Nutrition (EPIC)–Potsdam Study." The American journal of clinical nutrition 84.2 (2006): 427-433.
[26] Chawla, Nitesh V., Nathalie Japkowicz, and Aleksander Kotcz. "Editorial: special issue on learning from imbalanced data sets." ACM Sigkdd Explorations Newsletter 6.1 (2004): 1-6.
[27] Han, Hui, Wen-Yuan Wang, and Bing-Huan Mao. "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning." Advances in intelligent computing. Springer Berlin Heidelberg, 2005. 878-887.
[28] Seiffert, Chris, et al. "RUSBoost: A hybrid approach to alleviating class imbalance." Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on 40.1 (2010): 185-197.
[29] Kubat, Miroslav, and Stan Matwin. "Addressing the curse of imbalanced training sets: one-sided selection." ICML. Vol. 97. 1997.
[30] Z. Zheng, X. Wu, and R. Srihari. Feature selection for text categorization on imbalanced data. SIGKDD Explorations, 6(1):80-89, 2004.
[31] Galar, Mikel, et al. "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches." Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 42.4 (2012): 463-484.
[32] Herman, William H., et al. "Effectiveness and cost-effectiveness of diabetes prevention among adherent participants." The American journal of managed care 19.3 (2013): 194.