Business Analytics using Datamining - Project Report

Preventing exacerbation in respiratory patients by early prediction of exacerbation

Team 3 (MD Team)

Kun-Lin Tsai (100081021)
Yuan-Yu Zhang (102080593)
Chris Yen Chen Lo (f126776550)

Date: 13/01/2015

Executive summary

I. Business Goal

1. Customers: Public health care system (government public welfare department)
2. Problems: The expense on those patients is more than 3 million dollars per year (half patients compared to Diabetes, but double expense).
3. Goal: Establishing a prediction model for finding out exacerbators and preventing patients getting worse with following criteria: **a.** A standard protocol for predicting exacerbators **b.** Fast and efficient **c.** Low cost: throat swab is a cheap way for sampling patients DNA. **d.** Less measurements: this model is optimized with essential predictive variables **e.** High accuracy: correct prediction

II. Data mining goal

1. Goals: Selecting useful variables for creating predictive model: **a.** Dimension Reduction (from many variables) **b.** With comparatively few predictors **c.** With good AUC (correct prediction)

III. Data profile

1. Resource: competition from crowdanalytix (MODELING: Predict Exacerbation in patients with Respiratory Diseases)
2. Dataset: a. 330 columns numerical medical derived data b. 1000 columns nominal genetic data c. 4000 rows (patients) d. Supervised datamining with 1/0 (excerabator or not) e. 300 variables after data cleaning

IV. Methods

1. Data partition with withdraw method to increase data amounts: (training: validation: test) = (0.5:0.3:0.3)
2. Grouping all variables into two part: genetic data (nominal) and medical derived data (numerical). For genetic data: Gene score is produced based on whether patients are exacerbators or not. For medical derived data: important numerical variables is found based on PCA
3. Logistic Regression is performed by integrating both parts (oversampling method is used). AUC is used to evaluate the performance of created models.

V. Results and conclusion

Predictors	Area under curve (AUC)
Only V238 (medical derived)	0.791
All medical derived data (w/o gene score)	0.819
Both all medical derived data and gene score	**0.830**

A flexible different predictors-based model is generated. Based on different variables we obtained, we can easily predict exacerbators with different accuracy.

Project report

i. Problem description

I. Business Goal:

1. Customers: Public health care system (Major: government public welfare department, minor: health care committee from community or university)
2. Purpose: For decreasing lethal rate and lowering social costs (by preventing high potential exacerbators getting worse and efficiently performing medication)
3. Problems: Respiratory diseases are a worldwide, contagious disease. It can be easily spread out by droplets and air. Researches show that 10~20% patients get worse in following 6 months. The expense on those patients is more than 3 million dollars per year (half patients compared to Diabetes, but double expense).
4. Goal: Establishing a prediction model for finding out exacerbators and preventing them getting worse with following criteria:

A. A standard protocol for predicting exacerbators: currently, there is only physical examination for patients, no standard protocol existed.

B. Fast and efficient: Only DNA exam and several non invasive exams should be performed. In clinical, those data can be easily obtained. All procedure can be done in 2 hours.

C. Low cost: throat swab is a cheap way for sampling patients DNA. It's no need to perform expensive diagnosis with high technique.

D. Fewer measurements: this model is optimized with essential predictive variables. That is, we can only use limited variables for prediction. More variables we need, more money we have to pay.

E. High accuracy: the performance of this model should be good. Correct prediction is absolutely needed to fulfill our business goals.

II. Data mining goal:

1. Purpose: Finding out predictive variables and creating a model for preventing exacerbation. Variables should be as few as possible (but still good for prediction).
2. Goals: selecting useful variables for creating predictive model which is fulfill following criteria:

A. Dimension Reduction (from many variables): We have to find out most representative variables from more than thousand ones.

B. With comparatively few predictors: More variables may increase the cost and limit the usage of this model (Some hospital may not cover all exams, which means not all variables can be obtained by all hospitals). We hope our model can be applied worldwide.

C. With good AUC (correct prediction): In clinical research, researchers care about "how correct the model can predict" Thus, we use AUC to evaluate our performance.

ii. Data description

I. Data information:

1. Resource: Data is downloaded from the website of a competition. The competition from crowdanalytix (MODELING: Predict Exacerbation in patients with Respiratory Diseases). Holdout sets are well prepared.
2. Dataset profile: We use validation data set to build our model. Validation data sets contain:

A. 330 columns numerical medical derived data

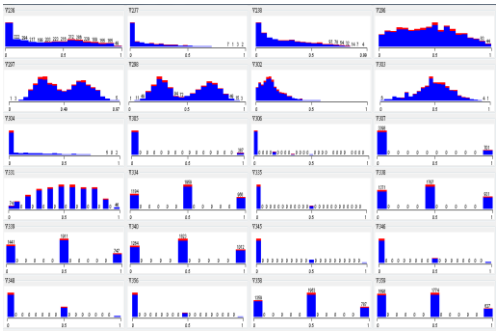
B. 1000 columns nominal genetic data

C. 4000 rows (patients)

D. Supervised data set with 1/0 (excerabator or not)

II. Data visualization and cleaning:

1. We focus on excerabation (1/0) and visualize all variables. Inside our data set, both nominal and numerical variables are existed. There is no such a variable that can perfectly separate excerabator.



sid	Exacebator	V3	V4	V29	V30	V31	V33	V36	V37	V38	V39
1296	0	0	1	0.49505	0	0.340659	0	0.47619	0.288395	0.522788	0.241652
1725	0	0	0	0.50495	0	0.626374	0	0.428571	0.480195	0.489276	0.439683
2195	0	0	1	0.554455	0	0.549451	0	0.404762	0.573315	0.687668	0.411311
2712	0	0	0	0.346535	0	0.373626	0	0.333333	0.387769	0.542895	0.324455
2913	0	0	0	0.376238	0	0.395604	0	0.345238	0.350938	0.580429	0.274078
1269	0	0	1	0.29703	1	0.274725	1	0.470238	0.213343	0.288204	0.269253
2836	0	0	1	0.356436	0	0.274725	0	0.386905	0.313412	0.601877	0.232388
91	0	0	0	0.514851	0	0.461538	0	0.529762	0.323141	0.449062	0.308242
5571	0	0	1	0.356436	0	0.21978	0	0.369048	0.309937	0.620643	0.222158
2645	0	0	0	0.326733	1	0.21978	0	0.464286	0.295344	0.352547	0.328315
4059	0	0	1	0.336634	1	0.406593	0	0.279762	0.364837	0.379357	0.387956
1724	0	0	0	0.405941	0	0.230769	0	0.494048	0.496178	0.754692	0.318278
5122	0	0	0	0.39604	1	0.175824	0	0.345238	0.314802	0.400804	0.323876
5566	0	0	0	0.376238	0	0.461538	0	0.488095	0.52467	0.829759	0.307663
3626	0	0	1	0.415842	1	0.296703	0	0.345238	0.246699	0.352547	0.275429
5254	0	0	0	0.376238	1	0.373626	0	0.428571	0.259903	0.365952	0.283343

Left: Based on excerabator, we visualized all variables (red: 1, excerabator, blue: 0, health)
Right: We get clean data set after cleaning missing values

2. Data cleaning: there are lots of missing values, "NA" and blank. After we remove those missing values, still, we have about 300 variables. Thus, we decide to ignore those missing variables since we still have plenty ones.

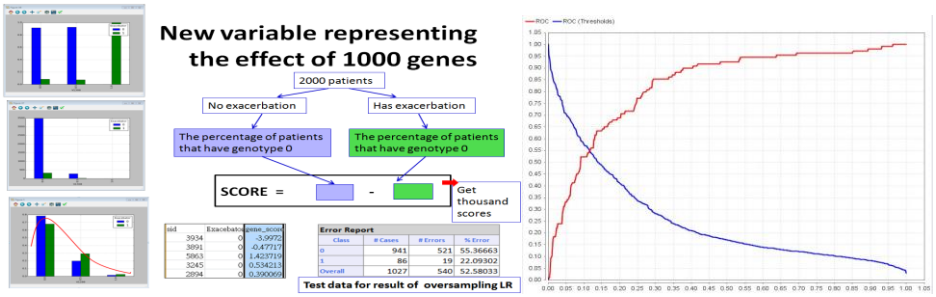
iii. Data preparation details

I. Data partition:

1. We generate hold outset by ourselves with the ratio training: validation: test =0.5:0.3:0.3. That is, we use replace method to increase the total amounts.
2. We use validation data to build our model, and test our model by testing data set.

II. Data mining steps

1. Separate numerical and nominal variables: we decide to use different approaches when dealing with variables because genetic data reflect different diseases progression than medical derived data.
2. For genetic data: Gene score is produced based on whether patients are exacerbators or not by determining the percentage for each variable.



Left: Gene score is generated for nominal data
Right: Our best model AUC value is 0.830 (red line)

3. For medical derived data: Important numerical variables are found based on PCA. PC1 and PC2 are chosen because they are considerably important. By ignoring other variables, we can generate reduced, simplified and easier model. We also analyze the relation between different variables and choose high potential predictors.
4. Combining both parts: Logistic Regression is performed by integrating both parts. That is, we transform nominal variables (genetic data) into a gene score (numerical). Then, we choose numerical variables from PC1 and PC2 and generate a nonlinear model based on logistic regression. In total, we generate three different models based on different combination (flexible to user). We use oversampling method (excerabator 1:0 = 1:1).
5. Performance: AUC is used to evaluate the performance of created models followed by most of clinical researches. Because for most clinical researchers, they prefer choosing solution with highest accuracy (correct prediction) to benefit majority of patients.

iv. Data mining solution:

In total, we produce three models based on different variables combinations. All of them have high predicting power. The AUC values are as following:

I. Only V238 from medical derived data we chosen : 0.791
II. All medical derived data we chosen without genetic data: 0.819
III. Both all medical derived data we chosen and genetic data: 0.830

v. Conclusions (advantages and limitations)

I. We generate a respiratory diseases model with high predicting power which can be applied into different field (social health care, community etc.). Also, it could be the first one written dictates for excerabator prediction.
II. It's a flexible predicting model which is suitable for different level medical diagnosis (different input variable can be chosen)
III. The ideas for generating this model correspond to biomedical meaning for diseases. We consider about Single nucleotide polymorphism (SNP) and different disease progression and use them to generate predicting model.
IV. Limitation: Wrong prediction is possible. To reduced misclassification, we need some knowledge about variables (knowledge based) and weighting for specific important ones or generating a tuning parameter for our model.