

Group 3 Presentation

LI Weihao 1155077142

LI Jinzhao 1155077016

Wang Yiqun 1155062115

Peng Zhichao 1155062015

Project 1: Fraud detection of insurance claims

October 16,2018

Outline

- ▶ Data Processing
- ▶ Selection Criteria
- ▶ Methodology
- ▶ Result Comparison & Best Model
- ▶ Limitations & Difficulties

Data processing

1. Time gap
2. Mapping to numeric values
3. Change to dummy variables
4. Create new features using PCA

Time gap

Variables concerned:

Year, Month, WeekOfMonth, DayOfWeek,
DayOfWeekClaimed, MonthClaimed, WeekOfMonthClaimed

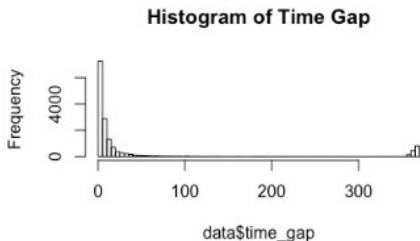
Goal:

Calculating the time gap between the claim and the accident

Time gap

Assumption:

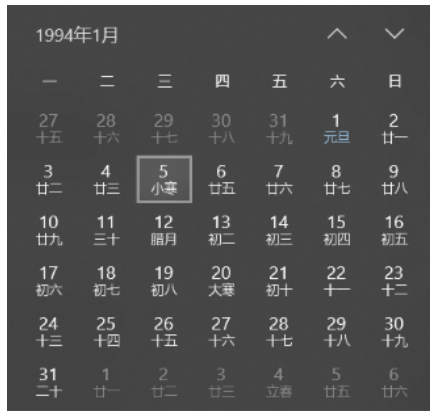
1. Assume claim and accident happened in the same year, but if claim happened before accident, assume that claim happened next year.



Time gap

Assumption:

2. Assume the first week of the month (WeekOfMonth) starts from the first day of the very month (instead of the first calendar week of the month).



1994年1月							^	v
一	二	三	四	五	六	日		
27 十五	28 十六	29 十七	30 十八	31 十九	1 元旦	2 廿一		
3 廿二	4 廿三	5 小寒	6 廿五	7 廿六	8 廿七	9 廿八		
10 廿九	11 三十	12 腊月	13 初二	14 初三	15 初四	16 初五		
17 初六	18 初七	19 初八	20 大寒	21 初十	22 十一	23 十二		
24 十三	25 十四	26 十五	27 十六	28 十七	29 十八	30 十九		
31 二十	1 廿一	2 廿二	3 廿三	4 立春	5 廿五	6 廿六		

Time gap

Method:

Date(MonthClaimed, DayOfWeekClaimed, DayOfWeek)

– *Date(Year, Month, WeekOfMonth, DayOfWeek)*

Mapping to numeric values

Variables concerned:

NumberOfCars, VehiclePrice, AgeOfVehicle, Days_Policy_Claim,
Days_Policy_Accident, PastNumberOfClaims,
NumberOfSupplements, AddressChange_Claim

Goal:

Converting the interval variables to some values so that they can be fitted into models.

Mapping to numeric values

Assumption:

Assume that for a certain interval variable, two values are similar if falling into the same interval. E.g. for NumberOfCars, “3 cars” and “4 cars” are similar as they fall into “3 to 4”

Method:

Taking the midpoint of the interval.

E.g. assign 3.5 to “3 to 4” in NumberOfCars

Change to dummy variable

Variables concerned:

Binary: Sex, AccidentArea, Fault, WitnessPresent,
PoliceReportFiled, AgentType

Categorical: MaritalStatus, Make, VehicleCategory, BasePolicy,
Deductible

Goal:

Change these variables to dummy variables

	MaritalStatus Divorced	MaritalStatus Married	MaritalStatus Single	MaritalStatus Widow
Divorced	0	0	0	0
Married	0	1	0	0
Single	0	0	1	0
Widow	0	0	0	1

Create new feature using PCA

Goal:

Help to improve the performance of the classifier and reduce the dimensionality of the data.

Method:

Perform PCA analysis to the dataset.

Random Under-Sampling

- ▶ Advantages

- ▶ It can help improve run time and storage problems by reducing the number of training data samples when the training data set is huge

- ▶ Disadvantages

- ▶ It can discard potentially useful information which could be important for building rule classifiers.
 - ▶ The sample chosen by random under sampling may be a biased sample. And it will not be an accurate representative of the population. Thereby, resulting in inaccurate results with the actual test data set.

Random Over-Sampling

- ▶ Advantages

- ▶ Unlike under sampling this method leads to no information loss.
- ▶ Outperform under sampling

- ▶ Disadvantages

- ▶ It increases the likelihood of overfitting since it replicates the minority class events.

Informed Over Sampling: Synthetic Minority

Over-sampling Technique

- ▶ Advantages

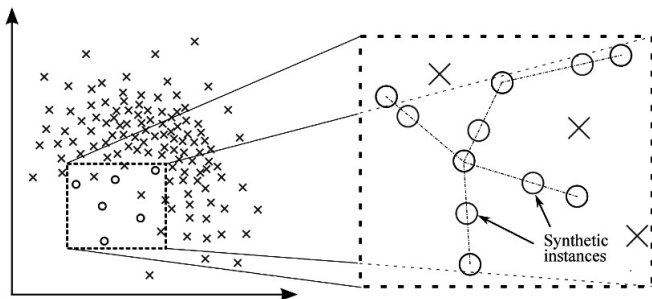
- ▶ Mitigates the problem of overfitting caused by random oversampling as synthetic examples are generated rather than replication of instances.
- ▶ No loss of useful information.

- ▶ Disadvantages

- ▶ While generating synthetic examples SMOTE does not take into consideration neighboring examples from other classes. This can result in increase in overlapping of classes and can introduce additional noise.
- ▶ SMOTE is not very effective for high dimensional data.

Informed Over Sampling: Synthetic Minority

Over-sampling Technique



Selection Criterion

- Precision is equal to the proportion of correctly raised alarms, as follows:

$$Pr = \frac{TP}{TP + FP}$$

- Recall is equal to the proportion of deviant signatures, which are correctly identified as such:

$$Re = \frac{TP}{TP + FN}$$

		Classified as	
Actual		Fraud	No fraud
	Fraud	TP-true positive	FN-false negative
	No fraud	FP-false positive	TN-true negative

Selection Criterion

- ▶ *F-measure* is a measure that calculates a harmonic mean between precision and recall, as follows:

$$F\text{-measure} = \frac{2 * Pr * Re}{Pr + Re}$$

- ▶ Use Recall and *F-measure* as final criterion and not use accuracy rate
 - ▶ Main purpose is detecting fraud cases
 - ▶ Common process
 - ▶ Select all potential fraud cases first
 - ▶ Classify selected cases artificially next
 - ▶ accuracy rate does not work under imbalanced situation
 - ▶ Accuracy rate is dominated by not fraud part
 - ▶ Extreme case: Given no fraud detection, accuracy rate=0.94

Results and Finding

		no change			PCA			
F_score		claims2_us	claims2_os	claims2_smote	claims2_pca_us	claims2_pca_os	claims2_pca_smote_os	
Supervised	Logistic Regression		0.2144	0.2098	0.2311	0.2119	0.2110	0.2120
	Classification Tree		0.2338	0.2171	0.2378	0.2081	0.1939	0.1940
	Random Forest		0.2415144(10/60)		0.2651934	0.2372159(10/27)		0.2577777
	KNN		0.1862	0.1872	0.1833	0.1801	0.1776	0.1707
	SVM		0.2109	0.2172	0.2167	0.2188	0.2180	0.2160
	Neural Network (5)		0.2323	0.2214	0.2149	0.2161	0.2298	0.1993
	Bagging	Bagging x Classification Tr	0.2386	0.2171	0.2513	0.2324	0.1958	0.2191
		Bagging x KNN	0.1723	0.1505	0.1747	0.1732	0.1483	0.1733
		Bagging x SVM	0.2110	0.2096	0.2115	0.2109	0.2112	0.2112
	Boosting	Boosting x Classification T	0.2138	0.2213	0.2254	0.1992	0.2170	0.2112
Boosting x KNN		0.1663	0.1250	0.1626	0.1578	0.1507	0.1683	
Boosting x SVM		0.2085	0.2090	0.2079	0.2256	0.2181	0.2389	
Unsupervised	K-mean clustering		0.2021	0.1865	0.1872	0.1955	0.2086	0.1900
	One Class SVM		0.0992	0.0998	0.1076	0.1011	0.1011	0.1048

Finding:

1. Most method give similar result, F score is about 0.22.
2. KNN and unsupervised one class SVM give bad result compared with others.
3. RF works better than others on average.

Results and Finding

- ▶ Bagging and boosting may be unchanged or even worsen the result(except the classification tree method using bagging).
- ▶ In the boosting, the weight decrease sharply, the first classifier dominate the result, which may make no difference with only one classifier.

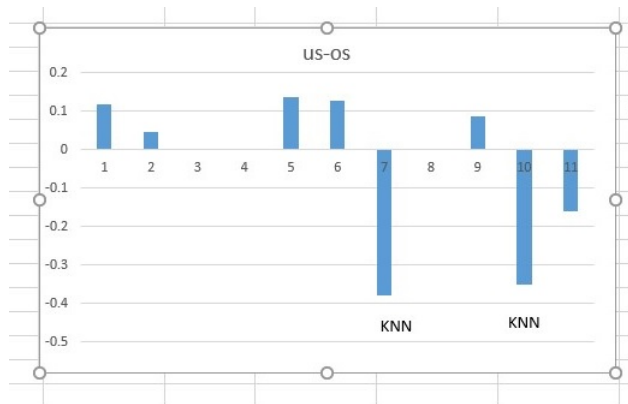
Results and Finding

Recall							
Supervised	Logistic Regression	0.7195	0.8374	0.8211	0.7115	0.8659	0.8618
	Classification Tree	0.9228	0.9675	0.8943	0.6545	0.7398	0.7195
	Random Forest	0.752		0.5854	0.6789		0.7073
	KNN	0.6237	0.6245	0.7284	0.58	0.5166	0.7538
	SVM	0.9187	0.9187	0.9431	0.8618	0.7967	0.9472
	Neural Network (5)	0.7846	0.9228	0.7927	0.7967	0.8171	0.7195
	Bagging	Bagging x Classification Tree	0.8415	0.9675	0.7967	0.5772	0.7846
		Bagging x KNN	0.5772	0.1992	0.561	0.5691	0.1992
		Bagging x SVM	0.9024	0.9024	0.9106	0.9268	0.9309
	Boosting	Boosting x Classification Tree	0.6318	0.7195	0.4932	0.5813	0.4715
		Boosting x KNN	0.5708	0.2194	0.5166	0.5448	0.2248
		Boosting x SVM	0.7834	0.6234	0.6543	0.7558	0.5317
		K-mean clustering	0.3862	0.3943	0.374	0.3699	0.374
Unsupervised	One Class SVM	0.4837	0.4878	0.4512	0.4837	0.4919	0.4472

Finding:

- ▶ Except the classification tree give worse result, PCA or not give similar result.
- ▶ SVM and Classification tree give higher recall compared with others.

Results and Finding



Finding:

- For the recall part, except the KNN, when the fraud case increase, most of classifier tend to have higher recall.

Result and Finding

Table 1: Best VS Ordinary

Result	0	1
0	3170	72
1	930	174
F	0.25778	
Recall	0.71	

Result	0	1
0	2558	34
1	1542	212
F	0.212	
Recall	0.86	

Observation:

Although the right result has higher recall, but it categorize many non-fraud to fraud, which make the F score lower.

Methodology

Random Forest is an extension of bagging with classification tree.

To understand RF, we need some knowledge about classification tree and bagging.

Methodology

Classification Tree

$$\begin{aligned} IG(T, a) &= H(T) - H(H|a) \\ &= - \sum_{i=1}^J p_i \log_2 p_i - \sum_a p(a) \sum_{i=1}^J -Pr(i|a) \log_2 Pr(i|a) \end{aligned}$$

where $IG(T, a)$ stands for information gain, $H(T)$ stands for the entropy of the parent nodes and $H(H|a)$ is the weighted sum of the entropy of the children nodes.

Algorithm:

1. Calculated the information gain of each possible first split.
2. Select the best first split that provides the most information gain.
3. This process is repeated for each impure node until the tree is complete.

Methodology

Bagging - Bootstrap aggregating

The random forests method applies the general technique of bootstrap aggregating, or bagging, to tree learners.

Algorithm:

Given a training set $X = x_1, \dots, x_n$ with responses $Y = y_1, \dots, y_n$,

1. Repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples.
2. Result is given by taking the majority vote in the case of classification trees.

Random Forest

It differs from bagging in only one way: at each candidate split in the learning process, a random subset of the features is selected. This process is sometimes called "feature bagging".

Limitations and Problems

- ▶ Time gap cannot be calculated correctly in some cases
 - ▶ There exists a complete fifth week in a month.
- ▶ Certain variables should provide exact data.
 - ▶ NumberOfSupplements
 - ▶ Days_Policy_Claim
- ▶ Few Numerical variables
 - ▶ Numeric variables transferred from categorical variables may not express real information
 - ▶ Classifiers working with numeric variables does not perform well.
 - ▶ ANN
 - ▶ Knn
 - ▶ SVM