TEL AVIV UNIVERSITY
Department of Economics
Data Science Seminar
Lecturer: Dr. Dan Zeltzer

Comparison of machine learning models for the imbalanced classification problem of fraud detection in car insurance claims

Authors: Elad Golan & Dov Tuch

Email: eladgo10@gmail.com , dovboristuch@mail.tau.ac.il

Date: 01/05/2022

Abstract

Automobile Insurance Fraud is one of the main challenges for insurance companies, causing yearly substantial financial losses worldwide and impairing the speed and quality of services to insured customers. In recent years, applications of machine learning models for fraud detection have been studied to deal with this problem. Often fraud detection problem is a private case of an imbalanced classification problem. A classification problem that concerns many other areas, such as credit card, health insurance, financial statements, disease detection, etc. Traditional classification algorithms can be limited in their performance on highly unbalanced datasets. In this study, we focus on comparing 3 machine learning models - Random Forest, Support Vector Machine, Naïve Bayes; and the combination of those models with imbalance classification techniques - Synthetic Minority Oversampling Technique, Near Miss Undersampling, cost sensitivity learning and feature selection decomposition. We evaluate models' performance by a few relevant metrics – Area Under ROC curve (AUC-ROC), F-scoring, G-mean and Area Under Precision/Recall Curve (AUC-PR). Of the 15 methods and models combinations considered in this paper, Random Forest with Cost Sensitivity technique achieved the best performance metrics. We have shown that the test results of the chosen model can be used as a preliminary filter of identifying a claim fraud for insurance companies and a high confidence that a classified non-fraudulent claim is indeed so.

Introduction

Insurance fraud is a deliberate deception perpetrated against or by an insurance company or agent for the purpose of financial gain. Common frauds include "padding" or inflating claims; misrepresenting facts on an insurance application; submitting claims for injuries or damage that never occurred; and staging accidents.

As economists our mission is to maximize overall social welfare. The lack of genuine information about the insurance claim causes an increase in policy prices and extends the time of receiving compensation, which also affects honest policy holder.

In the case of public institutions, such as Social Security / Ministry of Defense, the unnecessary payments are taken from public funds.

In this study, we compare machine learning models and common practices for fraud detection classification problems on car insurance claims dataset. This is important because developing accurate discrimination systems will prevent fraud success and will identify fraud faster thus reducing the overall time to handle claims and save money. We have chosen to focus on car insurance claims, but our work can be applied for detection classification problems (with some adjustments) in other fields, such as health insurance, credit card payment etc.

Literature overview

1) *SVMs Modeling for Highly Imbalanced Classification [1]*

In this work, different approaches of SVM model for dealing with an imbalance sample problem have been applied, SVM–WEIGHT, SVM-SMOTH, SVM-RANDU and a new based SVM method was introduced, Granular Support Vector Machines- Repetitive Undersampling algorithm (GSVM-RU). SVM–WEIGHT implements cost-sensitive learning for SVM modeling, the basic idea is to assign a larger penalty value to FNs than FPs. SVM-SMOTH adopts the SMOTE algorithm to generate more pseudo positive samples and then builds an SVM on the oversampled dataset. SVM-RANDU randomly selects a few negative samples and then builds an SVM on the undersampled dataset. GSVM-RU splits the training set into multiple negative information granules, in each iteration SVM is modelled to extract Negative Local Support Vectors (NLSVs) then these NLSVs are removed from the original training set. An aggregation operation is then executed to selectively aggregate the samples in these negative information granules with all positive samples to complete the undersampling process. Finally, an SVM is modelled on the aggregated dataset for classification. GSVM-RU gains a smart choice of undersampling, instead of a random choice as is done in SVM-RANDU.

Seven highly imbalanced datasets are collected from related works. The authors compared the SVM models and the previous approaches of the related works. The performance comparison was performed by 4 different metrics, G-mean, AUC-ROC, F-Measures, AUC-RR, better suited to binary imbalanced classification problems than the other traditional metrics. The result shows the GSVM-RU outperforms the most of previous best approach and it is also an efficient method. In addition, the results show that SVM-WEIGHT is also highly effective, but less efficient, therefore it can be the first SVM modelling method of choice if the dataset is not exceptionally large.

The above study contributed to our personal understanding, by exposing us to problems that can arise in imbalanced data and introducing us to relevant methods for solving. In addition, the article reinforced our understanding that it is important to look at a correct metric, depending on the specific classification / prediction problem.

2) *Feature selection for high-dimensional imbalanced data [2]*

In this paper a discussion on the problems of using feature selection with highly imbalanced data is conducted. By showing that the traditional model-independent feature selection methods have a biased influence toward the majority class. For example, the calculation of Fisher scores will be mainly influenced by the majority class and therefore will lower the metric score in the end. To reduce the bias, a model-independent decomposition-based feature selection method is proposed.

The proposed method is divided into three phases. First, a decomposition of the majority class into smaller sub-classes, using a clustering algorithm that can produce clusters with balanced size (Expectation-Maximization, k-means-clustering, etc…). This makes the data multi-class balanced. In The second phase, a traditional feature selection is used to select the best-m features from the multiclass data. In the third phase the sub-classes are relabeled to the original major class and the original labeled data with the selected features is returned.

Multiple data sets were used from different domains: CNS, LYMPH, OVARY, NIPS each data set has 7129, 7129 ,6000 and 13,649 features with class imbalance ratios of 2:1, 58:19, 25:8, 4:1, respectively. The performance metrics of each data set with decomposition feature selection and regular feature selection were compared using Naive Bayes, SVM and decision trees with

Bayesian learning (LIBSVM and C4.5 algorithms respectively). The results showed that the performance of decomposition-based FS became better than traditional FS when the number of m features selected is m >45. This study illuminated the importance of understanding how imbalanced data can influence traditional methods that are not necessarily the models themselves but also the processes we do beforehand.

3) *Mining corporate annual reports for intelligent detection of financial statement fraud–A comparative study of machine learning methods. [3]*
A comparison of different machine learning models for predicting fraudulent financial statements. The original data consisted of 311 fraudulent statements gathered from US SEC in AAER. The data was manually balanced by getting 311 non-fraudulent reports from firms with corresponding market capitalization and industry membership. 30 stratified samples (explain this) have been created from the original data. Divided to 75% training and 25% test. On each data partition a correlation-based selection filter combined with a forward-selection was used to find subsets of low inter-correlation and high correlation with the class. This filter was used for two reasons: highly inter-correlated data (P < 0.05) and model-independence. The average number of selected samples was 7.87 with S.D of 0.73
Afterwards, on each sample, a comparison of the performances of 14 different learning techniques was applied using Accuracy, TP rate, TN rate, MC (combination of FP and FN rates), F-measure and AUC. The comparison of the metrics of each method was implemented by a statistical paired t-test. The results show that ensemble methods had the best performance in terms of correctly classifying fraudulent claims, the MC metric of fraudulent firms was significantly higher than non-fraudulent.
Bayesian belief networks and Decision Table/Naïve Bayes (DTNB) hybrid classifier provided the highest accuracy on non-fraudulent firms in terms of TN rate. This suggests that correct prediction of non-fraudulent firms is less complex, where the detection of fraudulent firms requires more complex (and less interpretable) machine learning methods.
In the discussion they proposed the possibility of designing loss functions where misclassified fraudulent firms get a much bigger penalty 1:2, 1:10, 1:20,

the rates: used in the study, recommended for regulators and investors, respectively [15]. We learned that NB and RF are good models for fraud detection, and we used those models in our comparison as well as cost-sensitive learning.

Data description and preprocessing

The Vehicle Insurance Claim Fraud Detection data was taken from Kaggle[1] contains 15,000 vehicle claims reports between 1994 to 1996. Each report contains 33 variables, 1 continues, 32 categorical (ordinal and nominal), see Table 1[2]. Only 6% of observations have been labeled as fraud, as can be seen in Figure 1, hence it is an imbalanced classification problem. As seen in Table 1 and 2, there are features with many categories and there is a noticeably substantial difference between the number of observations in each category under both fraud and non-fraud, thus there are groups with very few observations.

Table 1:
Variables description

Variable name	Description	Values
FraudFound_P	Indicates whether the claim was fraud (1) or not (0)	1,0
Sex	------	male, female
Age	Age of driver who made the accident	numeric
Fault	Categorization of who deemed fault	Policy Holder, Third Party
Marital Status	Marital status of claimant	Single, Married ,Widow Divorced
Driver rating	Not Specified which which is better	Ordinal from 1-4
VehiclePrice	ranges of vehicle prices	less than 20000, 20000 to 29000, 30000 to 39000, 40000 to 59000, 60000 to 69000, more than 69000
BasePolicy	type of insurance coverage	Liability, Collision, All Perils
Days_Policy_Claim	the number of days between when the policy was purchased and the claim filled	None, 8 to 15, 15 to 30, More than 30
PastNumberOfClaims	previous number of claims filed by policy holder	None, ‘1’, 2 to 4, more than 4

[1] https://www.kaggle.com/datasets/shivamb/vehicle-claim-fraud-detection
[2] Main variables only.

PoliceReportFiled	indicates whether a police report was filed for the accident	yes, no
WitnessPresent	indicated whether a witness was present.	yes, no
VehicleCategory	Categorization of vehicle type	sport, sedan, utility

Figure 1:

Bar plot of Fraud and Not-Fraud frequency

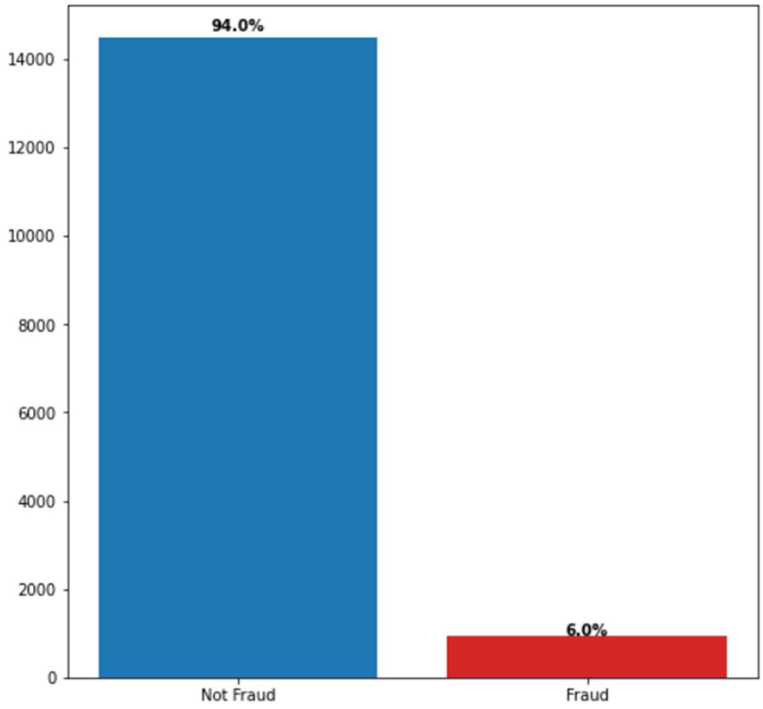


***Table 2**:*

Descriptive statistics

Variable name	category	Not fraud	Fraud
Sex	female	2315 (16%)	105 (11%)
	Male	12182 (84%)	818 (89%)
AccidentArea	Rural	1465 (10%)	133 (14%)
	Urban	13032 (90%)	790 (86%)
VehicleCategory	Sedan	8876 (61%)	795 (86%)
	Sport	5724 (36%)	84 (9%)
	Utillty	347 (3%)	44 (5%)
BasePolicy	All Perlis	3997 (28%)	452 (49%)
	Collision	5527 (38%)	435 (47%)
	Liability	4793 (34%)	36 (4%)
Fault	Policy Holder	10344 (71%)	886 (96%)
	Third Party	4153 (29%)	37 (4%)
VehiclePrice	Less than 20000	7658 (53%)	421 (46%)
	20000 to 29000	3358 (23%)	175 (19%)
	30000 to 39000	430 (3%)	31 (3%)
	40000 to 59000	83 (0.5%)	4 (0.5%)
	60000 to 69000	993 (7%)	103 (11%)
	More than 69000	1975 (13.5%)	189 (20.5%)
PastNumberOfClaimes	none	4013 (28%)	339 (37%)
	1	3351 (23%)	222 (24%)
	2 to 4	5191 (36%)	294 (32%)

	More than 4	1942 (13%)	68 (7%)
PoliceReportField	No	14085 (97%)	907 (98%)
	Yes	412 (3%)	16 (2%)
AgeOfVehicle	2 years	70 (0.5%)	3 (0.5%)
	3 years	139 (1%)	13 (1.5%)
	4 years	208 (1.5%)	21 (2%)
	5 years	1262 (9%)	95 (10%)
	6 years	3220 (22%)	228 (25%)
	7 years	5482 (38%)	325 (35%)
	More than 7	3775 (26%)	206 (22%)
	new	341 (2%)	32 (4%)

missing values: 320 of the observations were recorded with Age 0 and one observation was recorded with a 0 value for DayOfWeekClaimed and MonthClaimed. We assumed 0 indicates missing values, hence, to avoid dropping these observations, we replaced these values with the mean. We dropped PolicyType, PolicyNumber, AgeOfPolicyHolder and Make (car manufacturer) variable, because PolicyNumber is an ID of each Policy and PolicyType appears to be a concatenation of VehicleCategory and BasePolicy. AgeOfPolicyHolder and Age has correlation of 96%. Binary features replaced with zero one coding, ordinal feautres replaced with ordinal numbers or average of each category, using OneHot encoder for the rest categorical features. Afterwards, we split the data into train, validation, and test set in ratio of 70:20:10. Notice OneHot encoder application and the number of observations is relatively low in certain groups limitation may lead to sparse features matrix and cause issues in machine learning models like overfitting, inaccurate feature importance and high variance [4].

Imbalanced classification

The case when a data set is dominated by a major class or classes which have significantly more observations than the other rare/minor classes in the is known as class imbalance [1]. Imbalanced classifications pose a challenge for predictive modeling as most of the machine learning algorithms used for classification were designed around the assumption of an equal number of examples for each class [2]. It is possible that minority examples may be treated as noise by the learning model. This results in models that have poor predictive performance, specifically for the minority class. This is a problem because typically, the minority class is more important, therefore the problem is more sensitive to classification errors for the minority than the majority.

This problem becomes even more severe when the dimensionality of the data is high. Also known as the curse of dimensionality [3] which is out of the scope of this paper.

Table 3:
Confusion Matrix

		Real labeled	
		Positive (Fraud)	Negative (Not-Fraud)
predicted	Positive (Fraud)	TP	FP
	Negative (Non-Fraud)	FN	TN

To properly classify the minority class (fraud in our case) we must choose the right metric to optimize fraud detection. The common metrics are based on confusion matrix as shown at Table 3. When using balanced data, a common metric that is chosen is the accuracy measure. With highly skewed data distribution, the overall accuracy metric (1) is not sufficient anymore**.** Accuracy is a metric that is defined as the percentage of correctly classified samples across all classes [4]. It is useful when all classes are of equal importance, but in our case the important class is the minority (fraudulent claim). It is easy to get a high accuracy score by simply classifying all observations as the majority class [4].

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \qquad (1)$$

More suitable metrics are the Roc-Auc ,Area Under sensitivity/true positive rate (2) and specificity/true negative rate (3) Curve, F1-score (7), G-mean (4) and AUC-PR [5]. The tradeoff between sensitivity and specificity allows us to optimize our imbalanced classification problem relative to both classes, independent of class distributions. In fact, instead of maximizing the overall accuracy of the model we try to maximize true positive rate (true fraud classification) while considering the 1-specficty (false fraud classification) rate that we do not want to be too high. The Geometric Mean (G-Mean) is a metric that measures the balance between classification performances on both the majority and minority classes [6]. ROC-AUC can also indicate balanced classification ability between sensitivity and specificity, but unlike G-Mean by changing the threshold ROC-AUC can be used to choose between high sensitivity or low 1-specificity according to our specific problem [7].

$$sensitivity = TP/(TP + FN) \tag{2}$$

$$specificity = TN/(TN + FP) \tag{3}$$

$$G - Mean = \sqrt{sensitivity * specificity} \tag{4}$$

In our case, the goal is to identify frauds, but we do not want to classify observations as fraud at all costs, because this will hurt honest customers. For this reason, we consider F1-score as it combines precision and recall into one metric by calculating the harmonic mean between those two [8]. It is actually a special case of the more general function F-beta[3], where by changing beta, you can give more weight to recall or precision. Another reason for using F1-score in imbalance classification problem is that we get an indication that the accuracy is irrelevant, when F1 is equal to zero. Similarly to AUC-ROC, AUC-PR allows a trade-off between recall and precision using a threshold.

$$precision = TP/(TP + FP) \tag{5}$$

$$recall = TP/(TP + FN) \tag{6}$$

$$F - Measure = \frac{2 * precision * recall}{precision + recall} \tag{7}$$

After choosing the right model metrics for performance evaluation, some methods like resampling, cost-sensitivity learning and feature selection for imbalanced data (mentioned in the literature review) can help improve a model's performance.

For resampling methods, we use SMOTE: Synthetic Minority Oversampling Technique and NearMiss undersampling. SMOTE produces a new observation of the minority class by randomly picking a point from the minority class and computing the k-nearest neighbors for this point. Synthetic points are added between the chosen point and its neighbors. SMOTE's disadvantage is that it increases the likelihood of overfitting since it replicates the minority class [9]. NearMiss undersampling removes observations from the majority, when two points that belong to different classes are close to each other in the

[3] Multiply F1 by $\frac{1+\beta^2}{\beta^2}$

distribution. NearMiss can help improve the runtime of the model and solve memory problems by reducing the number of training data samples when the training data set is big, however it can lead to discarding useful information and getting a biased sample [10].

Cost-sensitive learning gives different misclassifications costs to each class. misclassification costs may be described by cost matrix *C,* with C(i, j) [11] being the cost

of predicting that an example belongs to class *i* when in fact it belongs to class *j*. In our case the cost of misclassifying fraud is greater than the cost of misclassifying non-fraud claim. It is often recommended to use the inverse rate (IR) in the data to weigh the cost ratio [12]. In our sample our IR is (1:16) cost to non-fraud and fraud misclassification, respectively. ratios of 1:10 or 1:20 were recommended also [13]. A disadvantage of cost-sensitive learning is that the true misclassification costs are often not known, and we must use heuristics. Another disadvantage is that it increases the time complexity of the model [16].

For imbalanced feature selection we will use the decomposition-based feature selection previously discussed [maamar number], in the second phase we will use kmean clusters and mutual information with Kbest.

As a result of what we explained above, using standard machine learning models will lead to poor results in imbalance classification problems, therefore we will combine the metrics and methods we presented along with the models we chose to use.

We apply 3 machine learning models - Random Forest, Support Vector Machine, Naïve Bayes using sklearn python package. The models' and methods' hyper-parameters were chosen by heuristics.

Experimental Results

Table 4 summarizes the results for all models and methods. The best results between the models with applied methods are marked with bold font. The first row shows the results of a baseline RF model with no resampling or cost-sensitivity methods being used. The accuracy of the baseline models is the highest with 94.1% while the Roc-Auc, G-mean, F1, F2 Scores are the lowest of all other methods. The results show that RF-SMOTH had the highest F1-score, SVM-SMOTH with decomposition feature selection had the highest accuracy. RF-CS outperformed the remaining models in ROC-AUC, G-mean F2-score and Auc-PR. Comparing the classification metrics between “conventional” and decomposition-based feature selection, conventional NB-SMOTH, NB-Nearmiss SVM-CS, surpassed D-based by across almost all metrics. D-SVM-SMOTH had the highest accuracy of all models but underperformed SVM-SMOTH in all other metrics. D-SVM-NearMiss had higher accuracy than SVM-NearMiss but underperformed in all other metrics.

Table 4:

Models' performance

Model	Feature Selection	Accuracy	Roc-Auc	G-mean	F1-score	F2-score	Auc-PR
RF baseline	-	94.49 %	50%	0	0	0	52.76%
RF-SMOTH	-	75.11%	67.44%	66.89%	**20.66%**	33.83%	36.81%
RF–NearMiss	-	52.22%	66.13%	64.25%	15.87%	30.73%	45.78%
RF–CS	-	59.64%	**75.6%**	**73.44%**	20.34%	**38.35%**	**52.65%**
NB-SMOTH	Kbest-MI	62.23%	67.0%	66.79%	17.43%	32.01%	41.89%
NB-SMOTH	D-Kbest-MI	40.91%	64.02%	58.52%	14.37%	28.99%	49.18%
NB–NearMiss	Kbest-MI	56.73%	65.19%	59.26%	15.98%	30.25%	42.52%
NB-NearMiss	D-Kbest-MI	54.39%	60.08%	59.74%	13.84%	26.36%	38.01%
NB-CS	Kbest-MI	56.14%	72.36 %	70.03%	18.54%	35.47%	50.72%
NB-CS	D-Kbest-MI	39.48%	63.82%	57.66%	14.24%	28.84%	49.69%
SVM-SMOTH	Kbest-MI	72.93%	67.4%	67.11%	19.93%	33.48%	37.61 %
SVM-SMOTH	D-Kbest-MI	**78.35%**	55.31%	48.87%	13.02%	19.56%	20.83%
SVM–NearMiss	Kbest-MI	45.96%	60.33%	58.13%	13.48%	26.66%	42.58 %
SVM-NearMiss	D-Kbest-MI	53.74%	53.92%	53.92%	11.42%	21.69%	31.52 %
SVM-CS	Kbest-MI	60.39%	74.61%	72.88%	20.13%	37.75%	51.22%
SVM-CS	D-Kbest-MI	41.75%	63.36%	58.52%	14.21%	28.59%	48.03%

Figure 3 compares the F1-score, precision and recall of each model-method on the validation sample. SVM-SMOTH and RF-SMOTH have the highest precision (12.53%, 11.91%). RF-CS, SVM-CS, NB-CS with the highest recall (93.5%,90.5%,90.5%).

Figure 3:

Bar plot comparison of models' F1-scoring, precision, recall

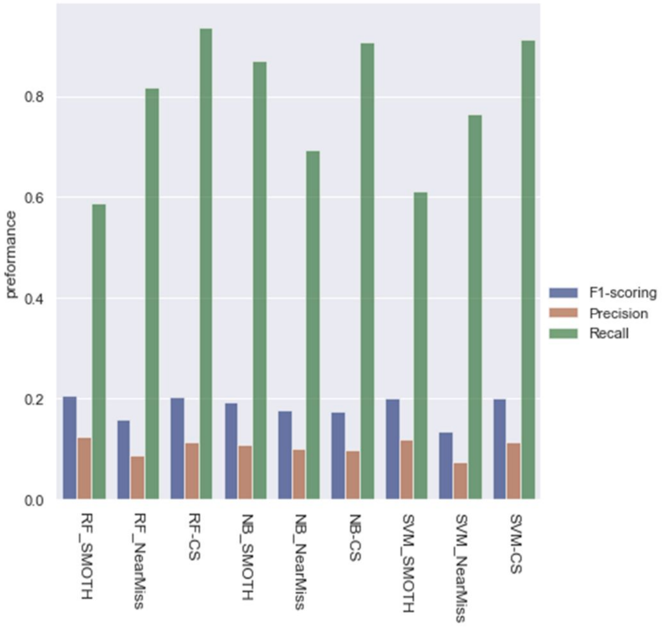
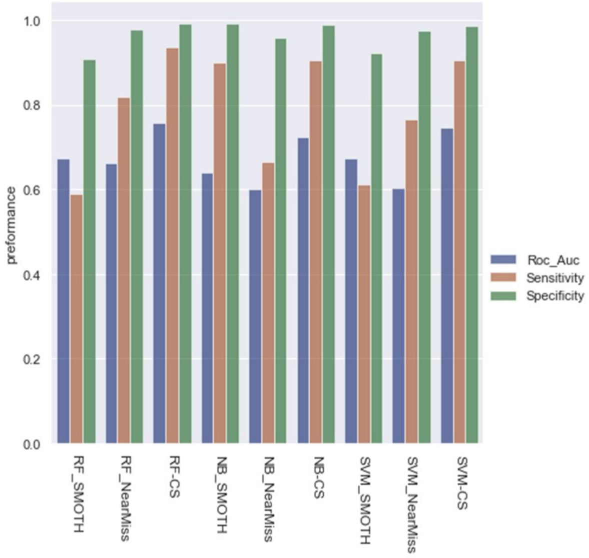


Figure 4 compares the Roc-Auc, Sensitivity and Specificity of each model-method on the validation sample. Roc-AUC and Sensitivity (recall) were discussed previously.

RF-CS, NB-SMOTH and NB-CS have the highest specificity rates (99.1%,99%, 98.81%)

Figure 4:

Bar plot comparison of models' Roc_Auc, Sensitivity, Specificity.



The chosen model is *RF-CS* because it performed best in most metrics. The scores and confusion matrix of the test set are:

- Accuarcy: 62.52
- Roc_Auc: 79.57 %
- G-mean: 77.18 %
- F1-score: 23.75 %
- F2-score: 43.65 %
- AUC-PR: 56.23 %

Figure 5:

Confusion matrix of the chosen model RF-CS.

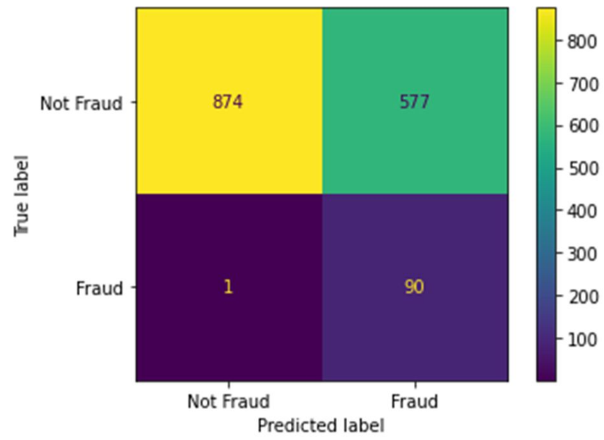
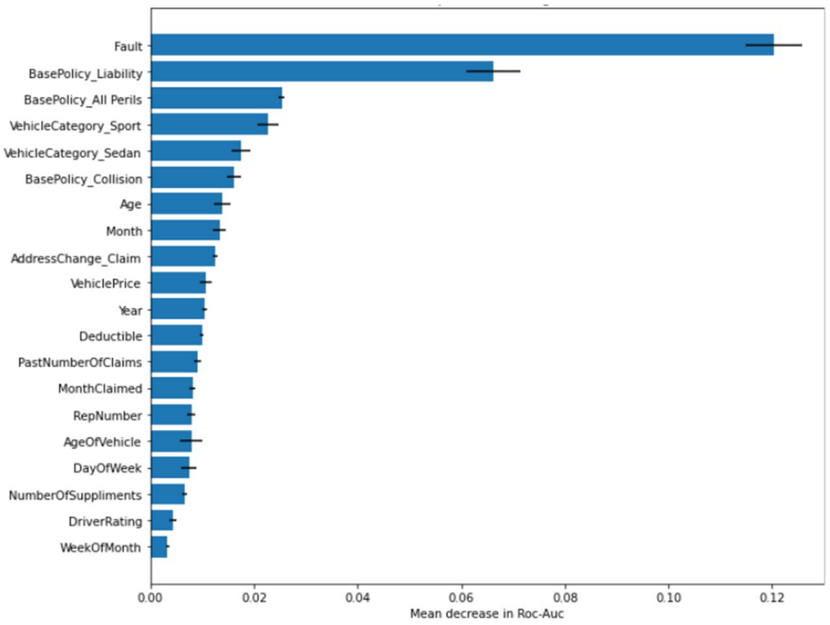


Figure 6 summarizes the results of feature importance using the mean decrease in Roc-auc of the chosen model (RF-CS), as can be seen from figure 2, the information about whose fault it was: Policy Holder or Third Party had the biggest influence on the decrease in roc auc by a large margin. The second most noticeable feature was whether the base policy was liability (damages to other people's property as well as medical costs).

Figure 6:

Feature Importance of the chosen model using Roc-Auc.

The black line represents Standard error.



Discussion

Elad Golan's Discussion:

We implement and compare machine learning methods on car insurance claims highly imbalanced dataset under 5 metrics (G-mean, AUC-ROC, F1-scoring, F2-scoring and AUC-PR). In previous works, imbalanced classification methods can be categorized into: oversampling, undersampling, cost sensitivity [1] [3] and imbalanced feature selection methods [2]. We establish our models based on the literature review.

The results show that the performances of three types of models are the lowest when we use under-sampling Near-Miss technique. A feasible cause for this result is the limitation of our data in terms of the relatively small groups in some categorial features and the disadvantage of using OneHot encoder (getting a Sparse feature matrix) as mentioned In Data description and preprocessing section. it may lead to discarding important samples and information when implementing under-sampling.

As seen in previous research [1] GSVM-RU and SVM-RANDU have good performance. Hence, we still recommend considering the implementation of under-sampling techniques, depending on the type and size of data available to the car insurance companies, besides it is the most efficient method.

The significant decrease in models' performances as a result of decomposition-based filter selection application may derive from the use K-means as part of the algorithm that performs not well when dataset consists of categorical, ordinal, and numeric variables that have not been normalized before using the method [19].

RF-CS (The chosen model) cannot be applied as an automated fraud detection algorithm because of low precision; however, it classifies right almost certainly as a fraudulent the test observation (few FN), therefore it can be implemented as a first filter for fraud detection. Only observations identified as fraud will be investigated further, thus reducing the number of claims that need to undergo credibility testing. As a result, insurance companies will save money, time and be able to invest more effort in identifying real potential fraud claims.

For further research, can be helpful examining additional methods that have been presented in the literature, improving the preprocessing, exploring of feature engineering applications and using methods such as cross validation to select hyper parameters in order to maximize each models' performance and for "fair" comparison.

Dov Tuch's Discussion:
As mentioned in the literature review, a comparison between learning models on a balanced data set and a discussion on the importance of feature selection was held [3]. Two of the other studies suggested different approaches when modeling highly imbalanced data. Resampling methods, cost-sensitive learning [1] and decomposition-based feature selection [2]. The present study was therefore designed to compare the combined effects of the best performing models from study [3] and the proposed methods in studies [1], [2]. The results of our study agree with the empirical evidence of study [3]. The cost sensitive ensemble method RF-CS performed best in terms of sensitivity (TP rate) and extremely high specificity rate as well as NB-SMOTH. Which means that if a claim is classified as non-fraudulent there is high probability that it is indeed so [3]. This could be useful for insurance companies to quickly handle deemed-honest claims, removing unnecessary processing time and allocating more resources to find the fraudulent claims. A possible explanation to RF-CS outperforming RF-SMOTH is that for the large data sets, cost-sensitive learning does often yield better results than oversampling [16] because the larger amount of training data makes it easier to estimate the class-membership probabilities more accurately.
The results of traditional feature selection beating decomposition based are concurrent with study [2]. It was shown there that only for features selection of k >= 45 decomposition outperforms traditional one. A possible explanation is that in study [2] they used data sets with at least 6,000 features before preprocessing. Our data had 33 features before preprocessing and we used feature selection of k <= 14. This suggests that we should not extrapolate those findings for data sets below 6,000. Another explanation is that by using OneHot encoding we made sparse features (no data points are missing, but most of them have zero value). It was shown [20] that k-means algorithm is not robust to sparse features and instead we should have used the entropy-weighted k-means that is better suited for this problem. Perhaps a future

study about the effectiveness of decomposition methods for data sets with a lower number of features should be held.

References

[1] Y. Tang, Y. Zhang, N. V. Chawla and S. Krasser, "SVMs Modeling for Highly Imbalanced Classification," in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 39, no. 1, pp. 281-288, Feb. 2009, doi: 10.1109/TSMCB.2008.2002909.

[2] J. Van Hulse, T. M. Khoshgoftaar, A. Napolitano and R. Wald, "Feature Selection with High-Dimensional Imbalanced Data," 2009 IEEE International Conference on Data Mining Workshops, 2009, pp. 507-514, doi: 10.1109/ICDMW.2009.35.

[3] Petr Hajek, Roberto Henriques, Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods, Knowledge-Based Systems, Volume 128, 2017, Pages 139-152, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2017.05.001.

[4] Arushi Prakash ,Working With Sparse Features In Machine Learning Models, www.kdnuggets.com/2021/01/sparse-features-machine-learning-models.html

[5] N. Japkowicz and S. Stephen, “The class imbalance problem: A systematic study,” Intelligent Data Analysis, vol. 6, no. 5, pp. 429–449, 2002.

[6] N. V. Chawla, N. Japkowicz, and A. Kotcz, “Editorial: special issue on learning from imbalanced data sets,” SIGKDD Explorations, vol. 6, no. 1, pp. 1–6, 2004.

[7] Pes B, Lai G. 2021. Cost-sensitive learning strategies for high-dimensional and imbalanced data: a comparative study. PeerJ Computer Science 7:e832.

[8] L. A. Jeni, J. F. Cohn and F. De La Torre, "Facing Imbalanced Data--Recommendations for the Use of Performance Metrics," 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, 2013, pp. 245-251, doi: 10.1109/ACII.2013.47.

[9] M. Kubat and S. Matwin, “Addressing the curse of imbalanced training sets: one-sided selection,” in Proc. of the 14th International Conference on Machine Learning (ICML1997), pp. 179–186, 1997.

[10] A. P. Bradley, “The use of the area under the ROC curve in the evaluation of machine learning algorithms.,” Pattern Recognition, vol. 30, no. 7, pp. 1145–1159, 1997.

[11] C. J. Van Rijsbergen, Information Retrieval, 2nd edition. London: Butterworths, 1979.

[9] Jason Brownlee, “SMOTE for Imbalanced Classification with Python”, January.2020 in Imbalanced Classification, machine learning mastery website blog.

[12] Jason Brownlee, “Undersampling Algorithms for Imbalanced Classification”, January.2020 in Imbalanced Classification, machine learning mastery website blog.

[13] Metacost: A general method for making classifiers cost-sensitive, P Domingos, Cited by 1970 s.

[14] H. He and E. A. Garcia, "Learning from Imbalanced Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, Sept. 2009, doi: 10.1109/TKDE.2008.239.

3979 paper citations.

[15] Abbasi, Ahmed, Conan Albrecht, Anthony Vance, and James Hansen. “MetaFraud: A Meta-Learning Framework for Detecting Financial Fraud.” MIS Quarterly 36, no. 4 (2012): 1293–1327. https://doi.org/10.2307/41703508.

[16] McCarthy, Kate & Zabar, Bibi & Weiss, Gary. (2005). Does Cost-Sensitive Learning Beat Sampling for Classifying Rare Classes. Learning. 10.1145/1089827.1089836. 210 citations

[17] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning: with Applications in R. New York: Springer, 2013.

[18] ebb, G. I.; Boughton, J.; Wang, Z. (2005). "Not So Naive Bayes: Aggregating One-Dependence Estimators". Machine Learning. 58 (1): 5–24. doi:10.1007/s10994-005-4258-6

[19] Ahmad, Amir & Dey, Lipika. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. Data & Knowledge Engineering. 63. 503-527. 10.1016/j.datak.2007.03.016.

[20] L. Jing, M. K. Ng and J. Z. Huang, "An Entropy Weighting k-Means Algorithm for Subspace Clustering of High-Dimensional Sparse Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 8, pp. 1026-1041, Aug. 2007, doi: 10.1109/TKDE.2007.1048.

Appendix

Link to GitHub repo of our work: https://github.com/Eladgo10/DS-Seminar-project-

Link to Jupyter notebook: https://github.com/Eladgo10/DS-Seminar-project-/blob/main/seminer-project.ipynb