Using Some Data Mining Approaches with Application on Insurance Data

Aya Shehata
Lecturer of Statistics, Statistics, Mathematics, and Insurance, Faculty of Business, Benha University

Dalia Sherif
Demonstrator, Statistics, Mathematics, and Insurance, Faculty of Business, Benha University

Zohdy Nofal
Professor of Statistics, Statistics, Mathematics, and Insurance, Faculty of Business, Benha University

استخدام بعض طرق التنقيب عن البيانات مع التطبيق على بيانات التأمين

د/ آية شحاته
مدرس الإحصاء بقسم الإحصاء والرياضة والتأمين
كلية التجارة – جامعة بنها

داليا شريف
معيد بقسم الإحصاء والرياضة والتأمين
كلية التجارة – جامعة بنها

أ.د/ زهدي نوفل
أستاذ الإحصاء بقسم الإحصاء والرياضة والتأمين
كلية التجارة – جامعة بنها

المستخلص:

سنتعلم في هذه الورقة عن استخراج البيانات وخوارزمياتها. ويدمج ميدان استخراج البيانات الأساليب الإحصائية التقليدية مع خوارزميات علم الحاسوب. من أجل استخلاص المعرفة من كميات هائلة من البيانات لتطبيقها في العلوم أو الحساب أو الصناعة. ونظرا لأن تقنيات استخراج البيانات متاحة الآن على نطاق واسع وتم توثيقها توثيقا دقيقا بلغات البرمجة الإحصائية مثل بايثون، فإن تطبيقها أصبح أيضا أيسر في السنوات الأخيرة. وتقترِح هذه الدراسة نماذج إحصائية ونماذج لتعدين البيانات تطبق على بيانات التأمين على السيارات. وقد استخدمت بعض تقنيات استخراج البيانات في هذا العمل لتجنب الإفراط في التهذيب، والتجهيز المسبق للبيانات، ومعالجة اختلال التوازن في البيانات، وتعديل الميزات الخارجية لإعطاء نتائج أفضل. ويجري في هذا العمل تنفيذ أربعة تصنيفات، بما في ذلك الانحدار اللوجستي Logistic) Regression)، (K nearest neighbor)، والغابات العشوائية (Random Forest)، ومشروع شجرة القرار (Decision Tree) . ويقيَّم أداء النماذج على أساس تقييم مصفوفة النتائج قائمة على أساس التقييم، والتذكير ، والدقة، وf1 -score، والاحكام ومعامل ارتباط ماثيوز (MCC)، وانحرافROC لحساب المنطقة الواقعة تحت المنحنى (AUC) بعد استخدام تقنية أخذ عينات فوقية من الأقلية التركيبية (SMOTE) ، وتظهر النتائج أن مستوى الدقة العشوائي بلغ أعلى ٨٣,٣٣ %، حيث بلغت قيم الإعادة ٧٦,٧٢ %، والقيم الدقيقة ٧٩,٤٦%، و٧٨,٠٧% f1 – score، وقيم معامل ارتباط ماثيوز (MCC) ٦٤,٦٦ %. وتتسم شجرة القرار بثاني أعلى درجة من الدقة تبلغ ٧٧,٥٦ %، حيث تبلغ قيم الإعادة ٦٩,٥٤ %، وقيم الدقة ٧١,٦٠%، و٧٠,٥٥% f1-score، وماثيوز معامل الارتباط (MCC) ٥٢,٤٤ %.

الكلمات الافتتاحية:

التنقيب عن البيانات، مصفوفة النتائج، تقنية أخذ عينات فوقية من الأقلية التركيبية (SMOTE).

Using Some Data Mining Approaches with Application on Insurance Data

Aya Shehata [a], Dalia Sherif [a], Zohdy Nofal [a]

[a] *Department of Statistics, Mathematics, and Insurance,*

Faculty of Business, Benha University

Abstract: In this paper we will learn about data mining and its algorithms. The field of data mining merges traditional statistical methods with computer science algorithms. In respect of extracting knowledge from massive amounts of data for application in science, computation, or industry. Because data mining techniques are now widely available and have been thoroughly documented in statistical programming languages like Python, applying them has also become more easier in recent years. This study proposes statistical and data mining models applying on car insurance data. Some data mining techniques have been used in our work to avert overfitting, preprocess the data, handle the imbalanced in the data and modify the outliers to give better results. Four classifiers, involving Logistic Regression, Random Forest, Decision Tree and K nearest neighbor are performed in our work. The performance of models is evaluated based confusion matrix, f1-score, recall, accuracy, Matthews correlation coefficient(MCC), precision and ROC curve to measure area under the curve (AUC) after using synthetic minority over-sampling technique (SMOTE). The results indicate the Random Frorest has the uppermost accuracy of 83.33%, with recall of 76.72%, precision values of 79.46%, f1-score of 78.07% and Matthews correlation coefficient values of 64.66%. Decision Tree has the second uppermost accuracy of 77.56%, with recall values of 69.54%, precision values of 71.60%, f1-score of 70.55%, and Matthews correlation coefficient values of 52.44%.

Keywords: Data mining, SMOTE, Confusion Matrix

1 Introduction

The application of statistical model-based methods for training computers to extract knowledge from massive datasets is the culmination of the data mining concept. The set of data depicts actual historical data. A data mining model will use this data set and various manners it finds to assort the data or forecast data in the future. It will do this by using a custom algorithm. The data is evaluated by a model's processing or analysis algorithm based on a mathematical formula that combines logarithmic, arithmetic, statistics, probability, and calculus. Numerous data mining models come with several algorithmic variants.

When it comes to accurately and economically predicting categorization difficulties, data mining is essential. Different data mining techniques, involving support vector classification, logistic regression, and random forest, have been applied to evaluate, define, and accurately forecast the data outcomes. This study uses statistical and data mining algorithms to demonstrate the best outcome for categorization prediction.

Our contributions for the suggested work:

- Statistical strategies have been employed to address feature outliers in a dataset in order to enhance the classifiers' performance.
- Using SMOTE technique to balance the data.
- In this research project, four classifiers and various data mining approaches are used to achieve the optimum outcome.
- Random Forest is the most accurate classifier out of the four, with an accuracy rate of 83.33%, recall, precision, and f1-score values of 76.72%, 79.46%, 78.07%, and 64.66%, respectively.

2 Related Works

Numerous studies have been conducted on the prediction-making process and the assessment of data mining model performance in categorization across various domains, including insurance science. An overview of a few of these publications that review work related to insurance science and data mining models is provided below:

Bhowmik (2011) predicted and presented fraud using decision tree-based algorithms and naïve Bayesian classifiers. He examined the confusion matrix-derived model performance metrics. Confusion matrix-derived performance

measurements include accuracy, recall, and precision. Because of its significant class skew, it is a trustworthy performance metric in many crucial fraud detection application domains.

Tao et al. (2012, October) developed a dual membership fuzzy support vector machine model for the purpose of identifying insurance fraud. Each sample is given a dual membership during the SVM training process based on the distance between the sample mean vector and itself; the dual membership that is assigned can be used to describe the imprecision of insurance fraud data. The empirical findings demonstrate that the fuzzy support vector machine model with dual membership outperforms other conventional insurance fraud identification models.

Senousy et al. (2019) offered a compelling and original model that explains how the Egyptian social insurance dataset is pre-processed using supervised learning methods. For the purpose of determining which of the three algorithms is more accurate and efficient, they have selected the Decision Tree, Naïve Bayes, and CN2 Rule Inducer algorithms. Following algorithm application, the outcomes demonstrated that the Decision Tree and CN2 Rule Inducer algorithms outperform the Naïve Bayes algorithm in predicting which individuals are covered by the social insurance program and which are not.

Abdelhadi et al. (2020) focused on cutting-edge statistical approaches and data mining algorithms that are the best approach for handling missing information in order to create a precise model to anticipate auto insurance claims using machine learning techniques. They developed the prediction model utilizing XGBoost, Decision Trees (DT), Naïve Bayes classifiers, Artificial Neural Networks (ANN), and Kaggle's public datasets, which comprise 30240 cases and 12 variables. The outcomes of the experiment demonstrated that the model produced appropriate results. Of the four models, the XGBoost model and Resolution Tree had the highest accuracy, with 92.53% and 92.22%, respectively.

3 Data Mining Models

Once the data has been pre-processed four different data mining classifiers—Decision trees, Random Forests, Logistic Regressions, and K-Nearest Neighbor—have been used to predict car insurance. Subsequently, the best classifier is compared against other models using performance indicators.

3.1. Logistic Regression [LR]

It is a classification technique based on the likelihood that a sample will belong to a class, even though it is referred to as regression. To forecast a categorical result, it makes use of both continuous and categorical

variables. Logistic regression yields an output that falls between 0 and 1, making it appropriate for tasks involving binary categorization. Binary logistic regression is the name given to the analysis when the dependent variable of choice has two categorical outcomes. However, the approach is known as multinomial logistic regression if the result variable has more than two levels. It's known as logit model. The function of the odds ratio can be described as follow:

$$odds_i = \frac{p_i}{1-p_i} = e^{\beta_0 x_0 + \beta_1 x_1 + \cdots + \beta_n x_n} \quad \text{Eq. (1)}$$

Where p_i is the probability of an event i,

$\beta_0 x_0 + \beta_1 x_1 + \cdots$, represents the regression model as β_0 the intercept, β_1 is the regression coefficient and x is the explanatory of variable.

3.2. K-Nearest Neighbor [KNN]

The KNN of a new set of data are determined, and the new set of data is classified based on the majority of its adjacent data. Despite the simplicity of this classifier, the test data classification is significantly influenced by the value of K. Although there are numerous approaches to determine the values for K, we can easily test this classifier several times using various values for K to determine which value yields the best outcome. The Euclidean distance between a test sample and the designated training samples serves as its foundation. The formula of the Euclidean is

$$d(x_i, x_l) = \sqrt{(x_{i1} - x_{l1})^2 + (x_{i2} - x_{l2})^2 + \cdots + (x_{ip} - x_{lp})^2} \quad \text{Eq. (2)}$$

Where x_l be an input sample with p features (x_{l1}, x_{l2}, …, x_{lp}), n be the total number of input samples (i=1,2,…,n) and p the total number of features.

3.3. Decision Tree [DT]

Decision Trees are primarily employed to solve classification problems. However, they can also be used to solve regression problems. This classifier is tree-structured, with internal nodes standing in for dataset attributes, branches for decision rules, and leaf nodes for each outcome as figure 1 show.

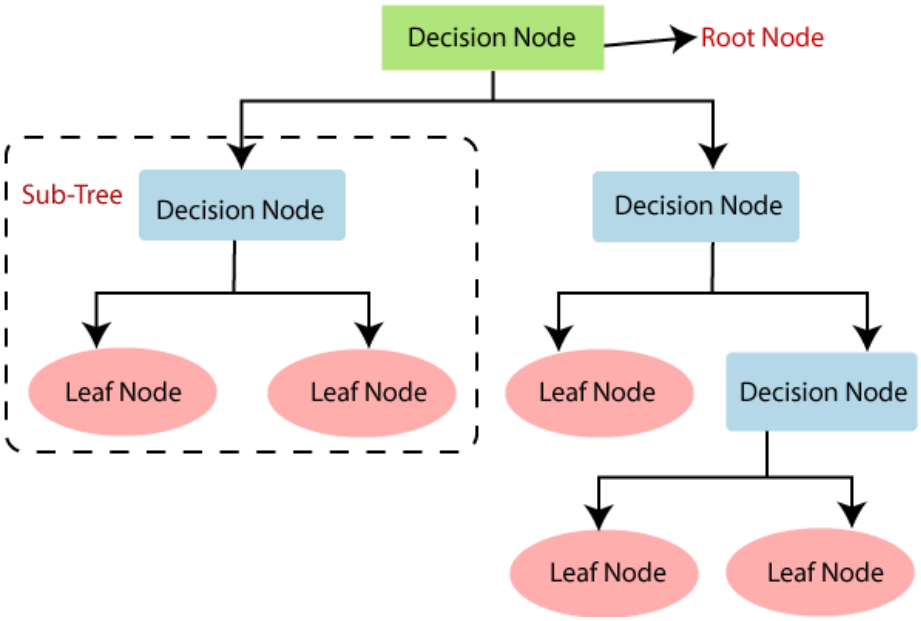

Figure 1 The general structure of the decision tree

Source: https://www.javatpoint.com/machine-learning-decision-treeclassification-algorithm

When the target is a classification outcome taking the values $0,1,...,i-1$, for a node j, representing a region D_j with observations N_j, P_{ji} is the proportion of class i observations in the node can be calculated as follows:

$$P_{ji} = \frac{1}{N_j}\sum_{y\in D_j} I\,(y = i) \qquad \text{Eq. (3)}$$

3.4. Random forest classifier (RF)

Random Forest is a popular machine learning algorithm that is a part of the supervised learning technique and may be used for both classification and regression issues in data mining. Its foundation is the concept of ensemble learning, which is the practice of merging several classifiers to improve the model's performance and solve a challenging issue. A Random Forest classifier is characterized by the presence of several decision trees on distinct dataset subsets, each of which is averaged to enhance the dataset's predictive accuracy. Random Forest forecasts the ultimate result based on the majority of votes from projections rather than relying solely on one decision tree. It does this by combining predictions from each tree.

4 Results and Discussion

25% and 75% of the data, respectively, have been used for testing and training the suggested work. Managing the data imbalance by SMOTE technique. A variety of data mining classifiers, including LR, KNN, RF, and DT, have been assessed. Using performance measurements such as the ROC curve to calculate the area under the curve (AUC), the confusion matrix, recall, precision, f1-score, accuracy, and Matthews correlation coefficient (MCC). the best data mining model among these four classifiers is found. The performance evaluation metrics for every classifier are covered in this section. Next, the performance indicators listed in Table 1 were used to evaluate each model classifier.

Table 1 The performance metrics are employed in this paper.

Performance Metrics	Mathematical Formula
Accuracy [*A*]	$A = \frac{TP + TN}{TP + FP + FN + TN}$
Precision [*P*]	$P = \frac{TP}{TP + FP}$
Recall [*R*]	$R = \frac{TP}{TP + FN}$
F1-score [*F*]	$F = \frac{2 \times R \times P}{R + P}$
Matthews Correlation Coefficient [*MCC*]	$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$

Each classifier's performance was assessed using the previously described performance evaluation measures; Table 2 displays the results. With recall values of 76.72%, precision values of 79.46%, f1-score values of 78.07%, and MCC values of 64.66%, it is evident that the Random Forest model yields the best accuracy, at 83.33%. With recall values of 69.54%, precision values of 71.60%, f1-score values of 70.55%, and MCC values of 52.44%, Decision Tree has the second-highest accuracy, at 77.56%.

Table 2 performance evaluation metrics for each classifier after SMOTE

Model	Accuracy	Recall	Precision	F1-score	MCC	AUC
RF	0.83334	0.76724	0.79464	0.78070	0.64662	0.82
DT	0.77556	0.69540	0.71598	0.70554	0.52442	0.76
KNN	0.75222	0.71552	0.66756	0.69071	0.48526	0.75
LR	0.75	0.74815	0.64331	0.69178	0.48740	0.75

5 Conclusion

Data mining classifiers are crucial for predicting the dataset class. Predicting classification problems using various data mining classifiers is covered in this paper. The suggested work is divided into multiple parts, such as loading the dataset, handling the imbalanced data, preparing the data, and evaluating the classifiers' performance. The findings indicate that the Random Forest model has the uppermost accuracy, at 83.33%, along with recall, precision, and f1-score values of 76.72%, 79.46, and 64.66% for MCC.

References

- Abdelhadi, S., Elbahnasy, K., & Abdelsalam, M. (2020). A proposed model to predict auto insurance claims using machine learning techniques. *Journal of Theoretical and Applied Information Technology*, *98*(22).
- Batra, B., & Kundra, S. (2019). Naïve classification approach for insurance fraud prediction. *International Journal of Engineering and Advanced Technology (IJEAT)*, *8*(5), 2378-2382.
- Bhowmik, R. (2011). Detecting auto insurance fraud by data mining techniques. *Journal of Emerging Trends in Computing and Information Sciences*, *2*(4), 156-162.
- Calderon, G. A. (2022). improving predictive performance within drivers insurance cross-border claims through a collaboration between ml and human experts.
- Giuseppe, B. (2018). Machine learning algorithms: popular algorithms for data science and machine learning.
- Healy, L. M. (2006). Logistic regression: An overview. *Eastern Michighan College of Technology*.
- https://www.javatpoint.com/machine-learning-decision-treeclassification-algorithm.
- Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, *4*(2), 1883.
- Senousy, Y., Hanna, W. K., Shehab, A., Riad, A. M., El-Bakry, H. M., & Elkhamisy, N. (2019). Egyptian Social Insurance Big Data Mining Using Supervised Learning Algorithms. *Rev. d'Intelligence Artif.*, *33*(5), 349-357.
- Tao, H., Zhixin, L., & Xiaodong, S. (2012, October). Insurance fraud identification research based on fuzzy support vector machine with dual membership. In *2012 international conference on information management, innovation management and industrial engineering* (Vol. 3, pp. 457-460). IEEE.