Insurance Sales Forecast Using Machine Learning Algorithms

Zuhal Kurt, Emrecan Varyok, Ege Baran Ayhan, Mehmet Turhan Bilgin, and Duygu Duru

Abstract Car accidents and the possible resulting loss of assets or life are issues for every car owner that must contend with some point in their driving life. Driving is an inherently dangerous act, even if it does not seem so at first, resulting in greater than 33,000 fatal vehi le crashes in USA in 2019 alone. However, the loss of life and possible damages can be reduced with the help of insurances. Insurance is an arrangement under which a person or agency receives financial security or reimbursement from an insurance provider in the form of a policy. Insurances help limit the losses of the customers when an undesirable event occurs, such as a car crash or a heart attack. Vehicle insurance provides customers monetary compensation after unfortunate accidents, provided they annually pay premium fees to the companies first. Our goal is to develop a machine learning algorithm that predicts customers who are interested in getting or renewing their vehicle insurance with the help of personal, vehicle, contact, and previous insurance data. The insurance sales forecast is helpful to companies, since they can then accordingly plan its communication strategy to reach out to those customers and optimize its business model and revenue, while also being beneficial to customers, who can go through the process and the aftermath of car accidents easier thanks to their monetary compensation. In this paper, the Health Insurance Cross-Sell Prediction dataset is used. The proposed model tries getting the value by training itself on a train and test dataset and will result in a categorical response feature based on the aforementioned data with the aid of well-known machine learning algorithms: k-nearest neighbors, random forest, support vector machines, Naive Bayes, and logistic regression.

Keywords Insurance prediction · Data analysis · Machine learning algorithm

Z. Kurt (✉) · M. T. Bilgin
Department of Computer Engineering, Atılım University, Ankara, Turkey
e-mail: zuhal.kurt@atilim.edu.tr

E. Varyok · E. B. Ayhan
Department of Automotive Engineering, Atılım University, Ankara, Turkey

D. Duru
Department of Chemical Engineering, Atılım University, Ankara, Turkey

A. K. Bashir et al. (eds.), *Proceedings of International Conference on Computing and Communication Networks*, Lecture Notes in Networks and Systems 394,
https://doi.org/10.1007/978-981-19-0604-6_3

1 Introduction

The objective of this paper is to classify customers based on their probability and desire to buy insurance based on their personal information, personal preferences, and the data of their owned cars. The machine learning (ML) algorithm tried predicting the likelihood of getting a positive or negative response of the customers getting an insurance by learning from labeled or tagged data and ending up with a classified response feature. This task needs multiple definitions to be fully understood. The first definition of this task comes with the options of classification or regression. Since the insurance prediction algorithm should classify the response and should categorize classes, the output class of this algorithm response is not considered continuous, so the task needs to be defined as a classification problem [1, 2].

The second decision of the insurance prediction algorithm is the selection of a supervised or unsupervised algorithm. All the data in this paper are labeled, and the data are separated as a train and test dataset, with the training dataset providing the 'response' feature as a variable. The model then makes predictions on the test class by assigning the categorical 'response' feature to the given dataset. Since supervised tasks predict classes, unsupervised tasks predict groups, and our task is a classification problem, all of these senses push our proposed algorithm to be a supervised learning algorithm [1, 2]. Finally, our model response feature is outputted as either a '0' or a '1'. So, all in all, with the previously discussed topics, our task can be defined as a binary classification problem with supervised learning.

The dataset used for this paper is, namely Health Insurance Cross-Sell Prediction, and gathered from the Kaggle website [3, 4]. The dataset contains three different *.csv files named as sample_submission, test, and train. The train data have 12 features and 381,109 records, the test data have 11 features and 127,038 records, and finally, the sample submission data have the same number of records with the same data, except it includes only the I.D. and 'response feature.' This final dataset is used to figure out if the final predictions are correct since 'response' is the target feature of this proposed model. There are no missing or mismatched values in this dataset. The training dataset is used and modified for this paper since during this study an uneven match of response features has been observed, so the train data were reduced to 165,582 records in order to even up the response rates and end up with a more efficient algorithm, which has succeeded.

With this goal, this paper presents an insurance prediction algorithm. Initially, we give a simple overview of the proposed machine learning algorithm, and then we conducted this model by using the Health Insurance Cross-Sell Prediction dataset to explain how this model can be used in practice. The remainder of the paper is coordinated as follows: The detailed representation of the proposed prediction model is explained in Sect. 2. The evaluation measurements that are used in this study are given in Sect. 3. The application of experiments on a real-world dataset and discussion of the experimental results are included in Sect. 4. Finally, the conclusion of this study is summarized.

2 Motivation—ML Algorithms

The main aim while working with the Health Insurance Cross-Sell Prediction dataset is to make the necessary classifications according to gathered information about customers and their vehicles to predict their response. Firstly, the proposed model is trained, and then it is tested based on a test dataset that has not been 'seen' from the model. Therefore, our task can be considered a supervised binary classification, the following ML algorithms are deemed the most appropriate, and they are k-nearest neighbors (k-NN), Naive Bayes, random forest (RF), logistic regression, and support vector machines (SVM) [5, 6].

2.1 Insurance Prediction Model

This paper predicted the likelihood of customers' interest in getting or renewing car insurance based on personal and vehicle data. To make this proposed model as efficient, a binary classification ML algorithm is considered as the best, the original data are preprocessed, appropriate features are selected, and the sample size is reduced to get the best prediction result. The best ML algorithms for this task can be considered such as k-NN, Naive Bayes, logistic regression, random forest, and SVM.

After the evaluations, the random forest algorithm is deemed the most successful, with 97% accuracy on the training dataset, and 91% results on the previously untested test dataset. This result shows a slight overfit, which the proposed model cannot solve completely, yet it can be able to reduce significantly compared to the original results thanks to preprocessing. For the most basic explanation, a random forest is ensembles of a group of decision trees. Random forest is a supervised learning method that outputs good results with various regression and classification problems.

The random forest algorithm also compensates for the common drawback of decision trees. The main problem of decision trees is that they tend to overfit on test on training data. In the random forest method, since every tree is randomized, their overfit condition also gets randomized. This randomization reduces overfit by averaging the results of decision trees in the forest.

Like some algorithms before, random forest works well with large and low-dimensional datasets. The random forest also fixes the overfitting problem that plagues decision trees and does not require the rescaling of data, unlike some other algorithms. The disadvantage of this algorithm is that it is a slow worker, needing much time to predict and train the data; however, our personal computers have high-end RAMs and CPUs, we have decided to use them anyway. The random forest test scheme inside the Orange3 environment is shown in Fig. 1.

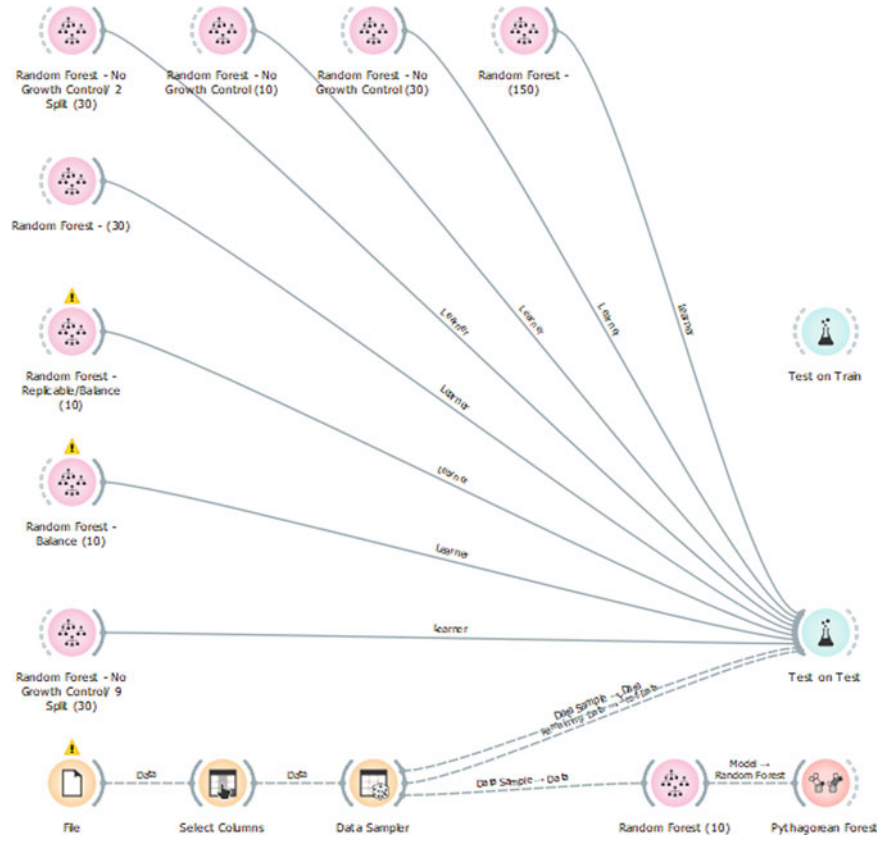

Fig. 1 Random forest test scheme inside Orange3 environment

2.2 Random Forest Parameters

In the Orange3, widget for the random forest algorithm has few options to modify (See in Fig. 2). Besides main parameters like the number of trees and the number of attributes, controlling the growth of every tree is possible.

The main parameter that changes results dramatically is the number of trees. The higher number of trees is tended to have better accuracies. However, the random forest is quite a source-hungry learning method. In about 150 trees, results were converged in a stable accuracy. Hence, the number of trees is set up to 150. Then, we changed other parameters around it to get different outputs. Initial results with different parameters show a ~10% difference between test and train accuracies. We tried to reduce overfitting by changing parameters, but it did not close the gap too much. Initial results were not good as assumptions; even 85.6% is not a terrible accuracy. The main problem is overfitting. Therefore, a simple preprocessing method is applied to the dataset, overfitting reduced as 5%. With such a little manipulation on

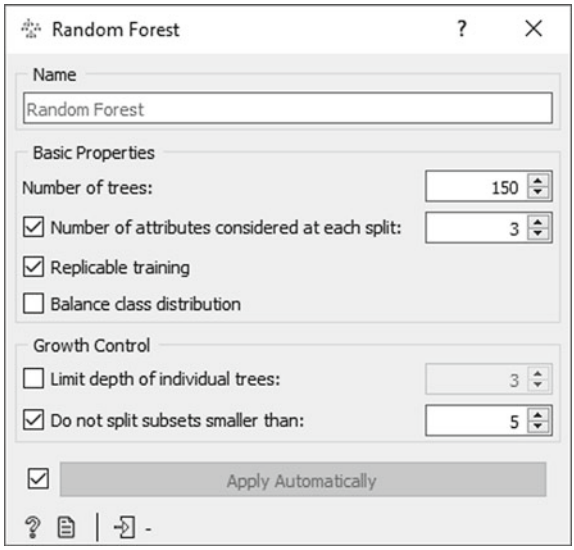

Fig. 2 Parameter screen of random forest in Orange3 environment

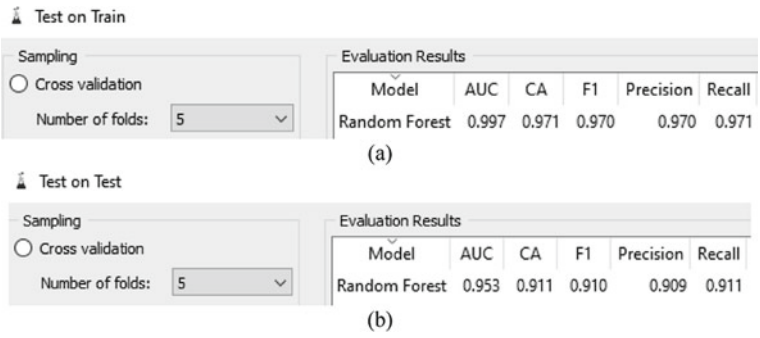

Fig. 3 **a** Test on train data results and **b** Test on test data results

the dataset, getting such improvement is satisfying for this study. After the preprocessing method is applied, the test on train data and test on test data results are shown in Fig. 3a, b.

3 Evaluation Measurements

3.1 Area Under Curve (AUC)

The area under the curve (AUC) is a performance metric for binary classifiers. By comparing the ROC curves with the AUC, it captures the extent to which the curve

is up in the northwest corner. A higher AUC is expected. A score of 0.5 is no better than random guessing, or a score of 0.9 can be considered a very good result, but a score of 0.9999 can be too good to be true and will indicate overfitting.

3.2 Classification Accuracy (CA)

Classification accuracy is represented as the ratio of the number of correct predictions to the total number of input samples.

$$\text{Classification Accuracy} = \frac{(\text{Number of Correct predictions})}{(\text{Total number of predictions made})} \tag{1}$$

3.3 Precision

Precision can be defined as the number of correct positive results divided by the number of positive results predicted by the classifier.

$$\text{Precision} = [\text{True Positives}/(\text{True Positives} + \text{False Positives})] \tag{3}$$

3.4 Recall

It can be defined as the ratio of the number of correct positive results to the number of relevant samples (all samples that should have been described as positive).

$$\text{Recall} = \left[\text{True Positives}/(\text{True Positives} + \text{False Negatives})\right] \tag{4}$$

3.5 F1-Score

F1-score can be referred as the harmonic mean between precision and recall. The range for F1-score is [0, 1]. This metric informs you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances).

High precision but lower recall gives you an extremely accurate, but it then misses many instances that are difficult to classify. Hence, the better performance of any

model can be expressed by the greater value of the F1-score. It can be mathematically represented as follows:

$$F1 = 2 * \left[1/(1/\text{precision}) + (1/\text{recall})\right] \tag{2}$$

4 Experimental Results and Datasets

Because of the research on related works, multiple similar types of research on car insurance prediction were found, two of which will be included here. For comparison, the results for the efficiency of our algorithms can be seen in Table 1. Effects of important features for insurance auto-renewal with classification ML algorithms are represented in [7]. In this research, the most successful models are random forest, gradient-lifting tree (GBDT), and lifting machine algorithm (LightGBM), with LightGBM producing the best result at 0.8045 AUC. However, our proposed method instead opted to use different methods, and due to the quality of our dataset, no missing data, independent features, and no necessary feature generation helped the proposed model to come out with better results with logistic regression. Our proposed model is resulting in a 0.924 AUC and is shown in Fig. 4. The data features of the dataset are mainly independent; hence, logistic regression works best with these types of data (See in Fig. 4). The experimental results of the proposed Health Insurance Cross-Sell Prediction model are given in Table 1.

Another related work, which used binary classification algorithms on a dataset gathered from a Brazilian car insurance company, tries to enhance the sales of a car insurance customer service [8]. Many methods similar to our approach used in this research, the eight common ML models are used, showed the best results using the random forest algorithms. The results of this research can be seen in Table 2. This research was also plagued with similar problems we have encountered during our study, with the binary classification algorithm incorrectly classifying 1 values as 0 due to noise and imbalanced data, with their final accuracy of the 1 values being 71%. The auto insurance model experimental results are given in Table 2. We have fixed this problem by removing 0-leaning data and reducing some unnecessary samples.

Table 1 Proposed Health Insurance Cross-Sell Prediction model performance

Measures (%)	Recall	Accuracy	Precision	F1-score	AUC
Logistic regression	0.919	0.919	0.916	0.917	0.958
Random forest	**0.920**	**0.920**	**0.917**	**0.918**	**0.961**
SVM	0.801	0.801	0.665	0.715	0.500
k-NN	0.919	0.919	0.916	0.917	0.955
Naïve Bayes	0.875	0.875	0.901	0.822	0.944

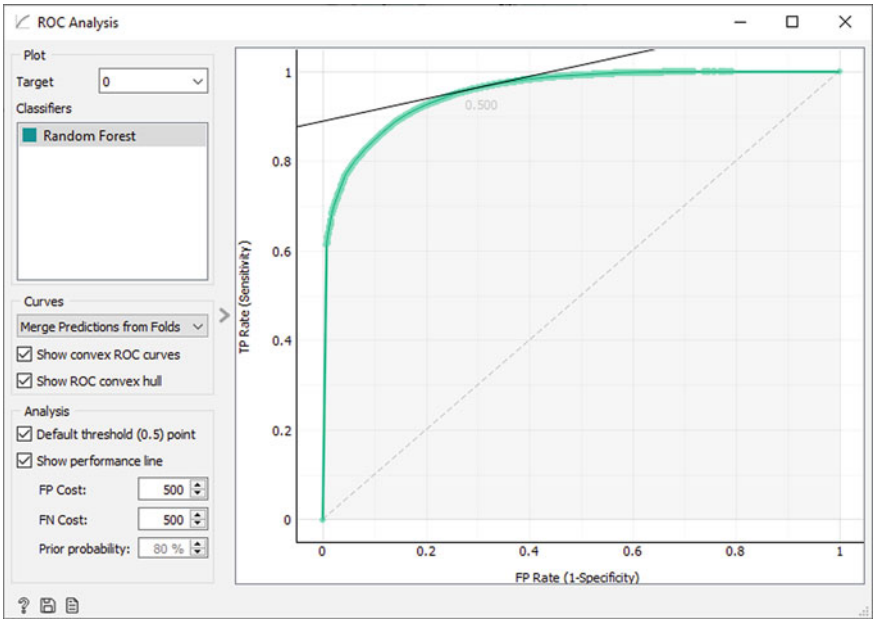

Fig. 4 Area under curve graph of random forest

Table 2 Auto insurance model performances [8]

Model	Accuracy	Precision	Recall	F1-score	AUC
RF	0.8677	0.9429	0.71	0.8101	0.84
C50	0.7913	0.7717	0.6743	0.7197	0.769
XGBoost	0.7067	0.6777	0.4994	0.575	0.671
J48	0.6994	0.6174	0.6399	0.6284	0.689
k-NN	0.6629	0.6167	0.4003	0.4855	0.628
LR	0.6192	0.55	0.2296	0.3239	0.615
Caret	0.6148	0.5601	0.1422	0.2268	0.534
Naïve Bayes	0.6056	0.6558	0.7273	0.6897	0.574

The performance is successful, with one of our best-proposed algorithms correctly classifying 85% of our 1 valued data, thanks to feature and sample reduction.

To obtain the general result of our proposed model, we also used tenfold cross-validation technique. The experimental results of the proposed model with tenfold cross-validation are shown in Fig. 5.

Fig. 5 Experimental results of the proposed model with ten-fold cross-validation

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
kNN-25	0.955	0.919	0.917	0.916	0.919
SVM c=10 g=10	0.500	0.801	0.715	0.665	0.801
Random forest-150	0.961	0.920	0.918	0.917	0.920
Naive Bayes	0.944	0.875	0.882	0.901	0.875
Logistic Regression 11	0.958	0.919	0.917	0.916	0.919

5 Conclusion

This paper aims to predict the likelihood of customers' interest in getting or renewing car insurance based on personal and vehicle data. It is a binary classification task, and we try finding a ML algorithm to solve this task. Furthermore, the original data are preprocessed, the important features are selected, and the sample size is reduced to get the best prediction result, in this paper. The most useful evaluation metrics are thought to be the confusion matrices on each result, as well the classification results on each model, and so these metrics are used. Overall, the algorithms worked significantly well on guessing the negative responses with almost all algorithms' final 0 response prediction rates being above 90%, yet the main shortcomings of the selected methods and algorithms were guessing the 1 response feature since the original data are heavily skewed toward the 0 responses. Even though that problem is eventually fixed with sample size reductions, the algorithms still have trouble with guessing 1 response rates. This problem can be solved with possible feature additions; however, any additional features can be created during this study, which is another shortcoming in and of itself.

The experimental result shows that the proposed model is solving this task efficiently. Furthermore, the experimental results demonstrate the superior performance of our proposed model by using five quality measurements: CA, recall, precision, AUC, and F1-score on the Health Insurance Cross-Sell Prediction dataset, as well as its flexibility to incorporate different information sources. To start with, from the companies' perspective, this study can lead to optimal insurance pricing, increasing company profits, and insurance customers. From the customer's perspective, since they are now more likely to be insured, they can have a monetary compensation in cases of vehicle accidents, resulting in better time and money management since insurance companies can help with the aftermath of the accident process. Finally, the drivers around the world can be safer around the roads since higher insurance having driver percentages can lead to safer roads and less reckless driving thanks to the drivers now being more careful because of their insurance driving guidelines, making the affected driver-used roads of this research safer in the long term.

References

1. S. Rawat, A. Rawat, D. Kumar, A.S. Sabitha, Application of machine learning and data visualization techniques for decision support in the insurance sector. Int. J. Inf. Manage. Data Insights **1**(2) (2021)
2. D.R. Gopagoni, P.V. Lakshmi, P. Siripurapu, Predicting the sales conversion rate of car insurance promotional calls, in *Rising Threats in Expert Applications and Solutions. Advances in Intelligent Systems and Computing*, vol. 1187, eds by V.S. Rathore, N. Dey, V. Piuri, R. Babo, Z. Polkowski, J.M.R.S. Tavares (Springer, Singapore, 2021)
3. IIHS HLDI, Fatality facts 2019 state by state: https://www.iihs.org/topics/fatality-statistics/detail/state-by-state#:~:text=There%20were%2033%2C244%20fatal%20motor,Columbia%20to%2025.4%20in%20Wyoming, Last accessed 31 June 2021 (2021)
4. K. Anmol, Health ınsurance cross sell prediction: https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction, Last accessed 31 June 2021 (2020)
5. B.D. Sommers, A.A. Gawande, K. Baicker, Health insurance coverage and health—what the recent evidence tells us. N. Engl. J. Med. **377**(6), 586–593 (2017)
6. Tutorials Point, KNN Algorithm—Finding nearest neighbors: https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_knn_algorithm_finding_nearest_neighbors.htm, Last accessed 31 June 2021
7. W.H. Dong, Research on the features of car ınsurance data based on machine learning. Procedia Comput. Sci. **166** (2020)
8. M. Hanafy, R. Ming, Machine learning approaches for auto insurance big data. Risks **9**(2), 42 (2021)