International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT228682

Vehicle Insurance Recommendation System

Ms. M. S. Sawalkar [1], Ruturaj Kumbhar [2], Krishna Jamkar [2], Mihir Mandlik [2], Harshawardhan Patil [2]
[*1] B.E.in Computer Engineering, JSPM Narhe Technical Campus, Pune, Maharashtra, India
[2]Assistant Professor, Department of Computer Engineering, JSPM Narhe Technical Campus, Pune, Maharashtra, India

Article Info

Publication Issue :
Volume 8, Issue 6
November-December-2022

Page Number : 584-589

Article History
Accepted: 05 Dec 2022
Published: 20 Dec 2022

ABSTRACT

Many automotive insurance providers are looking to improve their service for their customers, businesses are starting to adapt and implement machine learning and artificial intelligence methods of analysing data for performance, as a result giving better service for their customers from a better understanding of their needs. The main focus of this project therefore is targeted at automotive insurance providers looking to implement machine learning into their business, the project would also be beneficial to stakeholders and those who are looking to apply machine learning to improve their business. We propose a recommendation system built for a better customers experience, by suggesting them the most appropriate cover in time. The requirement for this system is to perform a more efficient up-selling than classic marketing campaigns.

Keywords: Automotive Insurance, Recommendation System, Machine Learning.

I. INTRODUCTION

In this report, we propose a recommendation system built for a better customers' experience, by suggesting them the most appropriate cover in time.

For most famous platforms, such as Amazon and Netflix, users must choose between hundreds or even thousands of products and tend to lose interest very quickly if they cannot make a decision.

Recommendation systems are then essential to give customers the best experience.

Recommender systems are machine learning algorithms typically employed to support marketing decisions, as they identify statistically validated associations between products and consumers.

These tools have been successfully adopted in many fields; however, not much has been done for the insurance industry.

We are constructing a recommendation system for car insurance, to allow agents to optimize up-selling performances, by selecting customers who are most likely to subscribe an additional cover.

The originality of our recommendation system is to be suited for the insurance context.

While traditional recommendation systems, designed for online platforms (e.g., e-commerce, videos), are constructed on huge datasets and aim to suggest the next best offer, insurance products have specific properties which imply that we must adopt a different approach.

Our recommendation system combines the XGBoost algorithm and the Apriori algorithm to choose which customer should be recommended and which cover to recommend, respectively.

Recently, the applicability of machine learning algorithms has become very popular in many different areas of knowledge leading to learn up-to-date advanced patterns from customers' behaviour and consequently target customers more accurately. In the context of recommendation systems, such algorithms generate automatically commercial opportunities suited to each customer.
Purpose: up-selling. Our goal is to support the agents that are and will continue to be the best advisers for customers, due to their experience and their knowledge of their portfolio.
In short, our tool helps them by automatically selecting from their large portfolios the customers most likely to augment their insurance coverage, in order to optimize up-selling campaigns for instance

II. THEORY

In order to make good decisions, it is necessary to possess ample amount of information. However, there are several examples showing that too much information is as bad as inadequate information; it is called information overload problem.

Recommender System has been introduced to solve this problem. It is very popular and useful concept in current digital era. It is an information filtering system that suggests products and services most relevant to the User.

Recommender System has been used widely for the products and services, intended for entertainment like music, books, and movies, online games, restaurants and completely based on user ratings.
Recommender systems are machine learning algorithms typically employed to support marketing decisions, as they identify statistically validated associations between products and consumers.
These tools have been successfully adopted in many fields; however, not much has been done for the insurance industry.

We are constructing a recommendation system for car insurance, to allow agents to optimize up-selling performances, by selecting customers who are most likely to subscribe an additional cover. The originality of our recommendation system is to be suited for the insurance context.

While traditional recommendation systems, designed for online platforms (e.g., e-commerce, videos), are constructed on huge datasets and aim to suggest the next best offer, insurance products have specific properties which imply that we must adopt a different approach.

Our recommendation system combines the XGBoost algorithm and the Apriori algorithm to choose which customer should be recommended and which cover to recommend, respectively.

III. DESIGN APPROACH

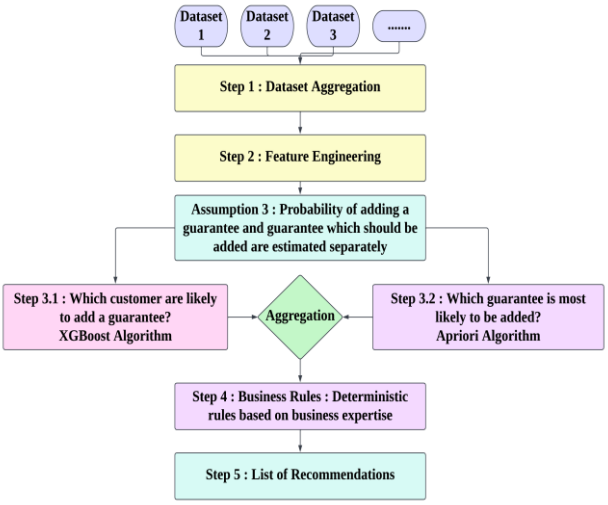

Figure 3: Dataflow Diagram

The model we propose the following approach to build the targeted recommendation system, illustrated by Figure 3.
After aggregating the different data sources (Step 1), we perform feature engineering (Step 2).
Assumption 3 is then made, illustrated by the separation of steps 3.1 and 3.2.

Step 1: Dataset aggregation : We built a unique dataset from multiple internal data sources. This dataset includes information about current customers' car policies (current cover, vehicle's characteristics, premium amounts), information about contacts between customers and the insurance company (phone calls, mails, etc.) and claims rate based on customers' history (in particular: claims not covered by their current covering).

Step 2: Feature engineering: Feature engineering allows us to build relevant features based on existing variables from the initial dataset. It could be an aggregation of several features, or a transformation from numeric to categorical feature. This step is in general based on knowledge of datasets and on intuition supported by experts from specific fields about what could be the most explanatory features.

Step 3.1: Step 3.1 takes as an input the customers' dataset x, defined by. Its outputs for each customer ui, his estimated probability to add a guarantee p i t .This probability is estimated by supervised learning using a label that represents whether a customer added a guarantee in the past or not, one year after the extraction date of features. Learning is performed on a training dataset, which is a random subset of the rows of x.

Step 3.2: Step 3.2 aims to predict which insurance cover/guarantee is most likely to be added, among the missing covers of the customers. This step answers the question: which additional insurance cover should we recommend? After testing several approaches, this step is performed by the Apriori algorithm.

Step 4: Business rules : Business rules consist in adding additional deterministic rules based on agent expertise and product knowledge. Given Property 1, this step avoids computing recommendations which are not usable in practice. For instance, some customers are already covered by recommended guarantee because they subscribed to an old version of car insurance product, whose guarantees were defined differently. Some simple business rules downstream of the model take into account these particularities.

Step 5: List of recommendations.

IV. RESULTS AND DISCUSSION

The learning algorithm which was trained on the data was the XGBoost classifier. This learning algorithm has mostly failed the first requirement as it is only capable of predicting the class '1' class once, and incorrectly classifying it four times. The class '0' predictions however have been classified significantly well, from these results however we can conclude the base model of XGBoost is incapable of classifying a claim or a no claim class, from this we can infer it will not be able to generalize on unseen data as it will overfit on the class '0' due to its large majority of predictions.

With regards to whether or not this initial XGBoost can produce good predictions for a good overall solution, it is possible but unlikely due to the lack of predicted class '1' classes. The Gini and roc score are fairly good results, however this initial XGBoost model is unlikely to generalize well on new unseen test data.

Table below displays the AUC values for the supervised learning algorithms.

Product	CTB	XGB	LGB
A	0.66	0.64	0.66
B	0.61	0.59	0.61
C	0.73	0.72	0.75
D	0.65	0.64	0.65
E	0.75	0.74	0.71
F	0.70	0.69	0.79
G	0.79	0.79	0.76
H	0.77	0.78	0.73

Table 4.1 : AUC values for supervised Learning Algorithms

ML models: AUC by product

Note: CTB = CAT Boost; LGB = lightGBM; XGB = XGBoost.

As expected, the predictive performance is always significantly higher compared to the canonical RS approaches, and to the GLMs, even if to a lesser extent. Furthermore, we note that there is no absolute winner among the chosen techniques, since the differences among them are small and could be attributed to suboptimal hyperparameter choices.

1) Validating the Chosen Classifier

Due to the fact that the Gini score and auc score are mathematically similar, in order to choose which model will be selected for the final classification model proposal, I will be comparing the auc and Gini scores relative to their visualizations on a receiver operating characteristic graph (roc).

The first figure represents the roc graph for the random forest classifier. Its curve is above the dotted red line which indicates 'random guessing', considering its auc score is 0.61 in this instance it is showing an unreliable score compared to its previously scored auc predictions.

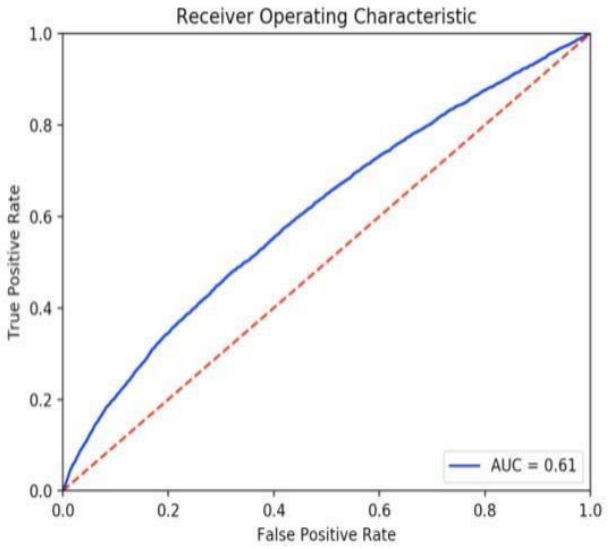

Figure 4.1 - A roc graph for the random forest classifier

As shown (below) by this XGBoost classifier, the auc results are better and therefore I decided to use XGBoost as the final classifier.

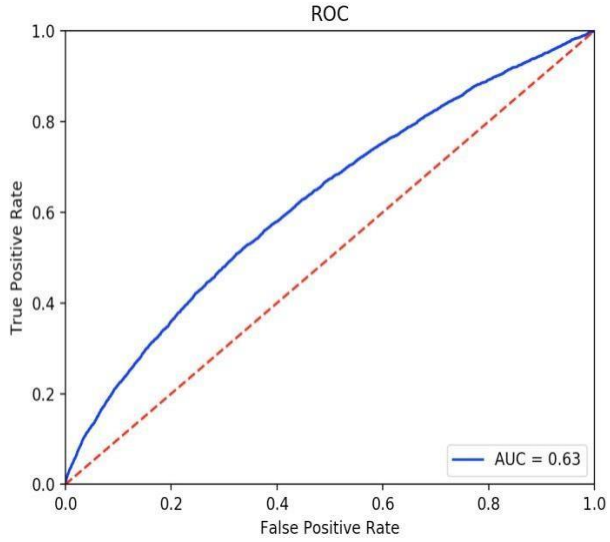

Figure 4.2 - A roc graph for the XGBoost classifier

V. CONCLUSION

The overall aims of this project were to produce a machine learning solution for predicting whether or not a new policy holder would produce an insurance

claim in the next year. This is in the hopes that it can make automotive insurance more accessible to more

drivers, through a solution that produces a more accurate prediction.

To make predictions it would have to effectively apply the feature information such as the customer's location or region, as well as their personal and vehicle information.

While it may have been slightly difficult to directly reference real-world features that were anonymized, this was possible to overcome by analyzing feature's importance through a correlation matrix and testing features normalized Gini coefficient score's individually on the dataset.

Using feature selection and training multiple classifiers we are able to produce a solution that is more complex than a simple classifier with parameters, and provide reliable claim predictions. It is important to factor in that the datasets provided were very noisy and imbalanced, and so the evaluation scores would not be possible to retrieve a nearly perfect score.

VI. FUTURE WORK

From carrying out this project we were intrigued by the results we had obtained from using various different learning algorithms.

Due to the short time given however, we did not explore in depth all the aspects of machine learning we would have liked to. In future we would like to train unsupervised and semi-supervised models, to determine whether or not their results could be used to improve the insurance claim predictions.

There are also certain aspects we would go further in depth for researching and selecting, such as the choice of categorical encoding methods like target encoding, as well as alternative data imputation methods such as the 'k nearest neighbour' algorithm which is more advanced for selecting more realistic replacement values, for values which are missing.

The addition of more models for testing and parameter tuning would also be a good way to improve on the project in future, because they can yield great results on various projects.

It would also be a good idea to train a live system (perhaps a batch system) which takes in batches of new customer information, trains the data and then produces predictions automatically within a reasonable time.

With a live system the data can thereby increase with more customer information and this could potentially lead to more accurate predictions if there is more data for the learning algorithms to work with.

Lastly due to the fact the data was anonymized we were unable to make discoveries about the features meaning in its real-world applications, we were only able to make assumptions with regards to common cover statistics.

If there had been more time we would have liked to try and specifically cross-reference feature values to certain statistics with more research, for example a feature that represents a car characteristics with regards to vehicle and producing claims.

VII. REFERENCES

[1]. Wu Shaomin. (2005). A scored AUC Metric for Classifier Evaluation and Selection. Accessed 18/05/2020, from http://dmip.webs.upv.es/ROCML2005/papers/wuCRC.pdf

[2]. Vladimir Kašćelan,Ljiljana Kašćelan,Milijana Novović Burić. A nonparametric data mining approach for risk prediction in car insurance: a case study from the Montenegrin market[J]. Economic Research-Ekonomska Istraživanja,2016,29(1) :545-558.

[3]. Chen M S , Hwang C P , Ho T Y , et al. Driving behaviors analysis based on feature selection

and statistical approach: a preliminary study[J]. The Journal of Supercomputing, 2018

[4]. Suyeon Kang,Jongwoo Song. Feature selection for continuous aggregate response and its application to auto insurance data[J].Expert Systems With Applications,2018(93):104-117

[5]. Alshamsi,Asma S. Predicting car insurance policies using random forest[C] 2014 10th International Conference on Innovations in Information Technology (INNOVATIONS). IEEE, 2014.

[6]. Xiaojun Ma,Jinglan Sha,Dehua Wang,Yuanbo Yu,Qian Yang,Xueqi Niu. Study on A Prediction of P2P Network Loan Default Based on the Machine Learning LightGBM and XGboost Algorithms according to Different High Dimensional Data Cleaning[J]. Electronic Commerce Research and Applications,2018(31):24-39.

[7]. Yanmei Jiang, Qingkai Bu. Supermarket Commodity Sales Forecast Based on Data Mining [J]. Hans Journal of Data Mining,2018,08(02):74-78.

[8]. Bobriakov, Igor. (2018). Top 10 Data Science Use Cases in Insurance. Accessed 12/02/2020, from https://medium.com/activewizards-machine-learning- company/top-10-data-science-use-cases-in-insurance-8cade8a13ee1

[9]. Sennaar, Kumba. (2020). How America's Top 4 Insurance Companies are Using Machine Learning. Accessed 20/04/2020, from https://emerj.com/ai-sectoroverviews/machine-learning-at-insurance-companies/

[10]. XGBoost. Accessed 07/02/2020, from https://xgboost.readthedocs.io/en/latest/index.html

[11]. Brownlee, Jason. (2020). Why One-Hot Encode Data In Machine Learning. Accessed 02/05/2020, from https://machinelearningmastery.com/why-one-hot- encode-data-in-machine-learning/

[12]. Friedman, N., Geiger, D. & Goldszmidt, M. (1997). Bayesian Network Classifiers. Accessed 20/04/2020, from https://doi.org/10.1023/A:1007465528199

[13]. Gusner, Penny. (2018). 13 things that affect your car insurance. Accessed 03/02/2020, from https://www.insure.com/car-insurance/car-insurancefactors.html#claims

[14]. Extremely Fast Gini Computation. Accessed 15/05/2020, from https://www.kaggle.com/cpmpml/extremely-fast-gini-computation Compare The Market. (2020).

[15]. Why is my car insurance so expensive? Accessed 2/04/2020, from https://www.comparethemarket.com/car-insurance/content/why-is-car-insurance-expensive

Cite this article as :

Ms. M. S. Sawalkar, Ruturaj Kumbhar, Krishna Jamkar, Mihir Mandlik, Harshawardhan Patil, "Vehicle Insurance Recommendation System", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 6, pp. 584-589, November-December 2022. Available at doi : https://doi.org/10.32628/CSEIT228682
Journal URL : https://ijsrcseit.com/CSEIT228682