iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VI **Month of publication:** June 2021

DOI: https://doi.org/10.22214/ijraset.2021.35642

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

Recommendations of Insurance Products based on User Behaviour

Sirisha Alamanda[1], Vaishnavi Devi Gujjari[2]
[1, 2]*Information Technology Department, Chaitanya Bharathi Institute of Technology*

Abstract: ***The remarkable increase in competition within the insurance sector has resulted in an overwhelming number of insurance products being available in the market. With rapid development of recommendation system, how to accurately predict Insurance policies using user lifestyle choices has become more and more important. The problem with the traditional systems is data sparseness. This paper proposes a recommender system to predict insurance products for new and existing customers. The main goal of the proposed system is to generate personalized recommendations based on the user lifestyle practices. By providing accurate personalised recommendations, the customer experience with the insurers can be improved.***
Keywords: Recommender systems, Random Forest, Logistic Regression, XGBoost, Decision Tree, Insurance Domain.

I. INTRODUCTION

A Recommender System is a selection guidance system that can help user filter out a subclass of information or predict user responses according to some preference criteria. For Example, Netflix's "*Recommended for you"* section, Amazon's "*Items you may like*" section etc.

Recommender systems provide insurance enterprises with insights on customer behaviours and demands, thereby serving as an important source of revenue generation. Insurers are seeking for a more efficient recommendation system to help them deliver the most suitable products to their customers. Most of the existing insurance recommendation systems are based on correlations between customers and products, ignoring the attempts hidden in customers' purchasing behaviours.

The proposed system delivers to offer products and options that are relevant to the customers, to help narrow down their set of choices to those most appropriate for them, and to improve the overall customer experience.

A. Applications

As one of the financial industries, the insurance industry is now facing a vast market and significant growth opportunities. The insurance company will generate a lot transaction data each day, thus forming a huge database. Recommending insurance products for customers accurately and efficiently can help to improve the competitiveness of insurance company.

Recommendation Systems have wide variety of applications. Insurance recommender provide users with more accurate and personalised recommendations as it not only considers basic portfolio data but also user habits that include whether user travels regularly, whether user uses public transportation and whether user exercises regularly etc.

The remarkable increase in competition within the insurance sector of the India has resulted in an overwhelming number of insurance products being available in the market. Having a recommender system in the Insurance industry will help the Insurers deliver more accurate results. The Insurance industry can grow additional revenues by incorporating Recommender Systems into existing applications and across enterprise solution as a means of monetizing product related data.

II. RELATED WORK

To our knowledge, there are not many documented instances for the insurance domain. Ref. [1] uses Bagging algorithm for data pre-processing where random sampling of data is done with multiple replacement. Random forest, an integrated learning method is used. The results obtained were compared from Nave-Bayes and Nearest-neighbours. The deployed system compared the error rates obtained when different models are used to recommend the insurance policies. The error rates obtained by random forest algorithm is less comparatively and has high feasibility in insurance product prediction [1].

Low Rank Matrix Factorization (LRMF) models are a widely used algorithm for recommender systems. Ref. [2] describes an approach to recommend insurance products using Bayesian networks which involves usage of Low rank matrix factorization. The deployed system suggests a tool to assist agents as a customer facing recommendation engine, and as a tool to assist the marketing department. The deployed recommender system aims to help the entire customer base by uniformly bringing coverage options to the attention of agents so that our customers are adequately covered for their needs [2].

Ref. [3] describes the usage of DCB algorithm driven Sankey diagram, that helps to reveal interpretable insights on the actual predictions of customers when purchasing insurance products. The attributes that are consider to recommend the insurance policies to the users are: age, gender, geographical region, purchasing plat- form, day-of-week, hour-of-day. The deployed model analyses the data both from insurance enterprise and cross media resources from the web and the algorithms chosen are applied for the prediction of results.

Behaviour Driven Visualization Recommendation (BDVR) [4] is a novel approach to visualization recommendation that monitors user behaviour for implicit signals of user intent to provide more effective recommendation. BDVR consists of two distinct phases 1. Pattern detection, 2. Visualization recommendation.

In the first phase, user behaviour is analysed dynamically to find semantically meaningful interaction patterns using a library of pattern definitions developed through observations of real-world visual analytic activity. In the second phase, BDVR algorithm uses the detected patterns to infer a user's intended visual task.

III.DESIGN CONSIDERATIONS

The following sections will describe an overview of the whole deployed system and the key design decisions made during the design process.

The deployed recommender system consists of the following modules, executed in-order:

A. The Recommendation Module

This is the core recommender model, it takes the input customer data. The dataset has multiple attributes which include hours worked per week, physical activity of the user, BMI, diabetic pedigree, if a person is a smoker or an alcoholic, travel frequency and also includes basic portfolio data like marital status, education, sex, age, children, work-class etc. The model then generates predictions for likely recommendations based on the conditions.

B. The Optimization Model

It takes as an input a set of possible recommendations from the module and prioritizes recommendations based on an optimization criteria predefined by the business. Possible choices for optimization strategies are: BMI, Diabetic pedigree, Travel frequency, number of children, physical activity, hours worked per week etc. The basic portfolio data like work class, marital status, sex are not considered in the policy allocation. Only the factors which describes user lifestyle habits and the user health data are considered in recommending the policies to the user.

C. The Business Rule Model

This module filters recommendations according to business rules to ensure that the final recommendations are appropriate for the customer. For example, it filters out a recommendation based on the conditions provided.

For example: A person with High BMI, working hours per week greater than 40, alcoholic and a smoker is most likely to be suggested a health Insurance policy and a person with Travel frequency greater than or equal to 3 is most likely to be suggested a travel insurance or pocket insurance based on the health conditions of the user i.e. the BMI, Diabetic pedigree of the user.

The proposed system is built as per the flow diagram shown in Fig.2.

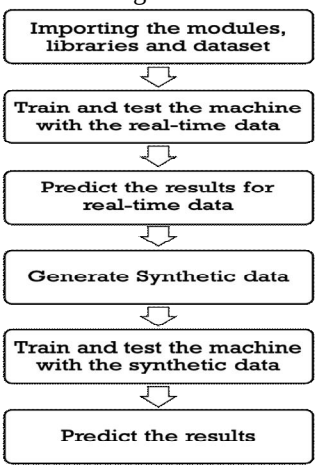


Fig. 1 Work flow of the proposed system

IV. RESULTS AND DISCUSSIONS

The testing of proposed system is carried out in 3 phases.

A. Data Source and Pre- processing

The data set used in the experiment is the real time dataset provided by US based Insurance company. The dataset includes attributes age, sex, education, relationship, work class, marital status, number of children, working hours per week, physical activity, BMI, diabetes pedigree, alcoholic, smoker and travel frequency.

TABLE I
SAMPLE REAL-TIME EMPLOYEE DATA

id	age	sex	Education	marital_ status	children	workclass	hours_ per_ week	physical_ activity	bmi	diabetes_ pedigree	alcoholic	smoker	travel
8046	39	1	10	1	0	8	40	23	30.1	0.398	1	1	3
7480	50	1	10	2	1	4	13	19	25.8	0.587	1	0	4
2122	38	1	14	6	3	6	40	0	30	0.484	1	0	1
8340	53	1	7	2	0	6	40	47	45.8	0.551	0	0	2
7533	28	0	10	2	0	5	40	0	29.6	0.254	1	0	4
8338	37	0	13	2	0	4	40	38	43.3	0.183	1	0	2
6454	49	0	5	4	1	14	16	30	34.6	0.529	1	0	4
3607	52	1	14	2	3	4	45	41	39.3	0.704	1	0	1
8057	31	0	13	1	2	5	50	0	35.4	0.388	1	0	1

The dataset is a real-time employee dataset with over 1300 entries. The proposed model is trained with the dataset and the predictions for the same are made.

The attributes age, number of working hours per week, number of children, Travel frequency, BMI, diabetes pedigree, alcoholic, smoker, physical activity are of high importance while education, sex, marital status are of low importance.

A Label Encoder is used to assign the labels as per the pre-defined conditions that are defined using the domain knowledge.

For example, a user with high BMI and unhealthy lifestyle habits like smoking and working more than 30 hours per week is more likely to be suggested a Health Insurance rather than Life Insurance or General Insurance.

The policies considered for the recommendation module which are encoded using the label encoder are as follows.

0 - General Insurance
1 - Life Insurance
2 - Health Insurance
3 - Pocket Insurance
4 - Retirement Insurance
5 - Travel Insurance.

As the dataset is of small size, in order to obtain more accurate results, the size of the dataset should be increased. In order to do that, a synthetic dataset needs to be generated with the attributes similar to the existing attributes. One of the main reasons for generating synthetic data is that when a model is trained with a large amount of data, there is a chance that the model gives more accurate results.

B. Synthetic Data Generation

Synthetic data is information that's artificially manufactured rather than generated by real-world events. Synthetic data is created algorithmically, and it is used as a stand-in for test datasets of production or operational data, to validate mathematical models and, increasingly, to train machine learning models. Synthetic data can be generated using multiple tools in Python. There are multiple libraries that can be used to generate artificial/synthetic data. The data generated can be numerical, binary, or categorical.

The real-time data is analysed and used in generating the synthetic datashown in Fig.2. The minimum and the maximum values of the attributes are set using the domain knowledge and the dataset is generated. The generated synthetic data preserves the mean of the attributes i.e. real-time data and the synthetic data will have the same mean. The technique used to generate synthetic data for the current dataset involves usage of Numpy, Scikit-learn and Random value generators. The proposed model has used standardized data given in the existing dataset to generate synthetic values. The generation of synthetic data has increased the prediction accuracy in all the algorithms used.

```
df
```

	id	age	sex	education	marital_status	children	workclass	hours_per_week	physical_activity	bmi	diabetes_pedigree	alcoholic	smoker	travel
0	8046	39	1	10	1	0	8	40	23	30.1	0.398	1	1	3
1	7480	50	1	10	2	1	4	13	19	25.8	0.587	1	0	4
2	2122	38	1	14	6	3	6	40	0	30.0	0.484	1	0	1
3	8340	53	1	7	2	0	6	40	47	45.8	0.551	0	0	2
4	7533	28	0	10	2	0	5	40	0	29.6	0.254	1	0	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
41473	4511	27	1	14	1	2	3	50	16	0.0	0.970	1	0	1
41474	2015	27	1	10	1	0	6	16	13	40.6	0.205	0	0	4
41475	3115	30	0	14	2	3	6	35	38	30.0	0.831	0	0	2
41476	7286	30	1	7	2	0	4	40	0	34.2	0.137	0	1	1
41477	2554	33	0	14	1	0	4	50	0	42.2	0.299	0	1	3

41478 rows × 14 columns

Fig. 2 Synthetic dataset

C. Experimental Analysis

The experiment is carried out with different training and testing percentages i.e. the experiment is carried out with 70% training data and 20% testing data, 75% training data and 25% testing data, 80% training data and 20% testing data using several algorithms like Logistic regression(LR), Random forest(RF), XGBoost(XGB) and Decision tree(DT).

The results obtained from the Logistic regression(LR), Random forest(RF), XGBoost(XGB), Decision tree(DT) algorithms are compared and the model with the more accuracy is chosen as the final model for recommending the insurance products to the user. The accuracy generated by each of the algorithm when trained and tested on real data with the train-test percentages as 70 and 30 respectively are shown in Table II.

TABLE II
Accuracy on Real-Time Data Set

Model	Accuracy
Logistic Regression	78.358
Decision Tree	98.756
Random Forest	97.014
XGBoost	99.751

The accuracy generated by each of the algorithm when trained and tested on synthetic data with the train-test percentages as 70 and 30 are shown in Table III.

TABLE III
Accuracy on Synthetic Data

Model	Accuracy
Logistic Regression	79.974
Decision Tree	99.991
Random Forest	99.887
XGBoost	99.983

As it can be seen from Table II and III, there is an increase in the prediction accuracy of the models when trained with large amount of data (i.e. synthetic data).

The model when trained and tested with 70 and 30 percent respectively has the higher accuracy compared to the other train-test percentages. Hence, in the proposed system, the train-test percentages are finalised as 70 and 30 respectively.

The prediction accuracy of XGBoost is greater than those of Decision Tree, Random Forest and Logistic regression in the mentioned order. Therefore, the recommendation based on XGBoost algorithm is superior to the other algorithms.

V. CONCLUSION

Traditional recommendation algorithms encounter problems such as cold start and sparse data when recommending insurance policies. As rapid technological advances reshape the insurance landscape, carriers must become more customer-centric, enhance customer service, create better solutions for operational efficiency. The proposed system aims to offer insurance products and options that are relevant to customers, to help narrow down their set of choices to those most appropriate for them, and to improve the overall customer experience.

The recommendations made are done using the conditions that are defined using the domain knowledge. The proposed system considers relevant data which can be used to recommend a particular insurance product to the user. This technique also uses user personal in addition to customer portfolio data which will further help in delivering users with most accurate and personalised recommendations of Insurance Products.

In the near future, the system can be improved by considering more user characteristic and lifestyle data to give more accurate and personalised recommendations, more policies to be recommended to the user, hereditary problems for Health Insurance can be considered. Go beyond cross-sell and up-sell targets to “next best action” for any customer at any time. E.g., send a timely communication to the customer or grant an additional discount.

REFERENCES

[1] Yan Guo, Xiaonan Hu, Yu Zhou, Wenchuan Cheng, “Research on Recommendation of Insurance products based on Random Forest”, in 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI).

[2] MaleehaQazi, Glenn M. Fung, Katie J. Meissner, Eduardo R.Fontes, “An Insurance Recommendation System Using Bayesian networks”, in RecSys’17, Aug.27–31, 2017.

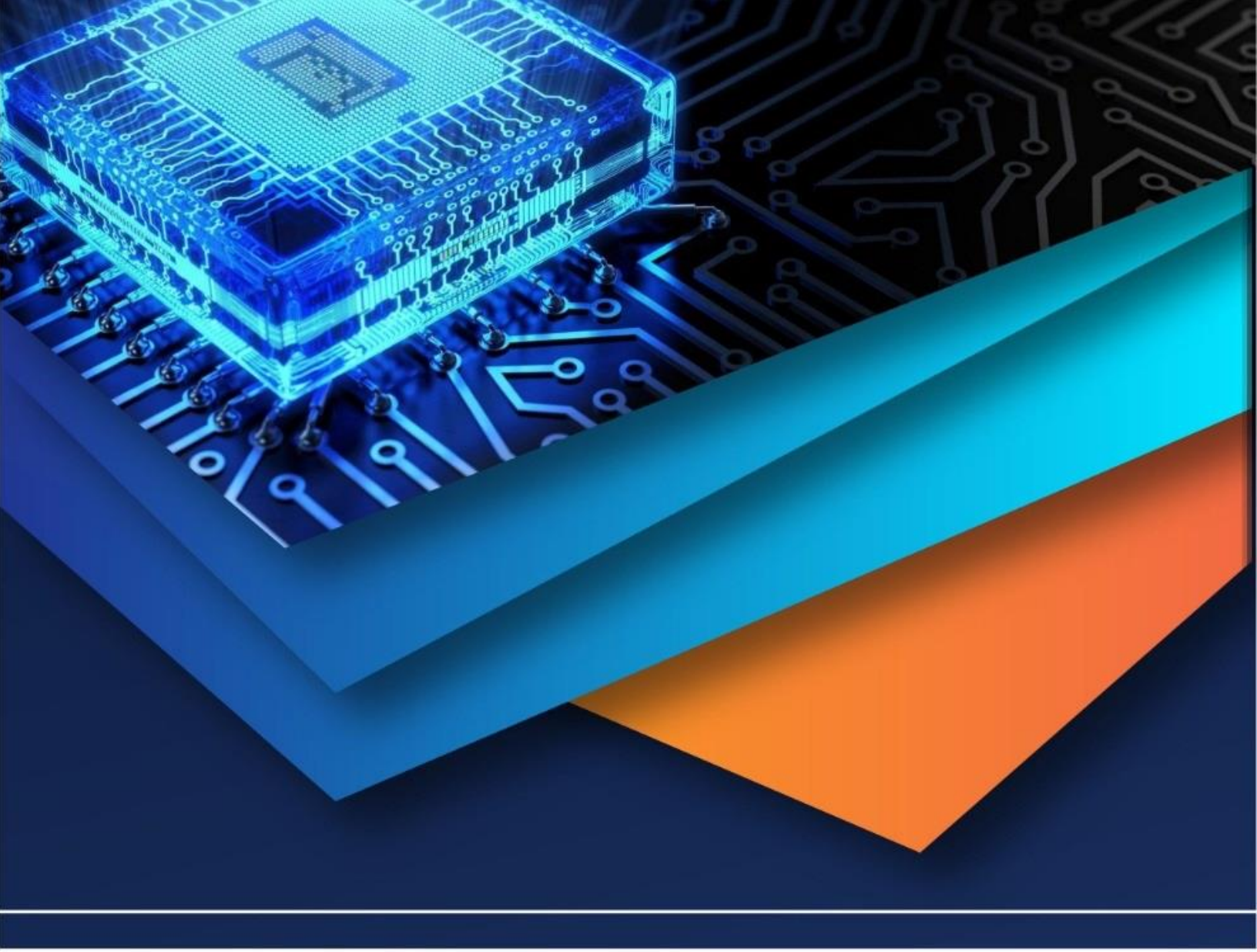
[3] Zhixiu Liu , Chengxi Zang , Kun Kuang , Hao Zou , Hu Zheng , Peng Cui “Causation-driven visualizations for insurance recommendation”, in 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW).

[4] David Gotz and Zhen Wen, “Behaviour-driven visualization recommendation”, in Proc. 14th Inter- national Conference on Intelligent User Interfaces, New York, USA, 2009, IUI ’09, ACM, pp. 315–324.

[5] Pakgohar A, Tabrizi R S, Khalili M, Esmaeili A. “The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach”, “Procedia Computer Science”, 3(none): 764-769, 2011.

[6] Rutkowski L, Jaworski M, Pietruczuk L, Duda P. “The CART decision tree for mining data streams”, Information Sciences, 2014.

[7] J. S. Breese, D. Heckerman, and C. Kadie. “Empirical analysis of predictive algorithms for collaborative filtering”, in Proc. Fourteenth Conference on Uncertainty in Artificial Intelligence, UAI’98, San Francisco, CA, USA, pp 43–52, 1998.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY