IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Predicting Customer Interest in Vehicle Insurance: A Study of Health Insurance Policyholders

[1]Dhruv Jore, [2]Shreyas Arora, [3]Amey Dubey, [4]Vikas Khare
[1]Undergraduate Student, [2]Undergraduate Student, [3]Undergraduate Student, [4]Associate Professor
[1]School of Technology and Management,
[1]SVKM's NMIMS Deemed-to-be University, Indore, India

Abstract: This research paper aims to predict customer interest in purchasing vehicle insurance from a health insurance company using data science techniques. The study uses data collected from health insurance holders of the company and employs data processing, analysis, and visualization techniques to prepare the data for machine learning models. Three models, namely logistic regression, Naive Bayes, and random forests were trained on the data, and their performance matrices were compared to determine the best-performing model. The study concludes that the random forests model outperformed the other two models in predicting customer interest in purchasing vehicle insurance. The results of this study have practical implications for health insurance companies seeking to expand their service offerings to customers. The study contributes to the growing field of data science in the insurance industry and highlights the potential benefits of using machine learning models to predict customer interests. The results of this research can aid insurance companies in developing targeted marketing strategies and improving customer satisfaction.

Index Terms **– Customer Interest, Machine Learning, Logistic Regression, Random Forests, Naive Bayes**

I. INTRODUCTION

The insurance industry is a highly competitive and dynamic sector that relies on understanding the behavior of its policyholders to optimize its business model and revenue streams. In today's market, insurance companies are constantly seeking ways to improve their sales and marketing strategies, and predictive modeling can provide valuable insights into customer behavior. One area of interest for insurance companies is the likelihood of a customer who has purchased a health insurance policy also being interested in purchasing a vehicle insurance policy.

This research paper explores this issue by building a predictive model using demographic information, information about the customer's vehicle, and information about their health insurance policy.

The goal of this study is to determine the factors that influence a customer's decision to purchase both types of insurance policies and to use this information to develop a tool that can be used by the insurance company to target its marketing and sales efforts more effectively. The data used for this study was provided by an insurance company that includes information on many policyholders who have purchased health insurance policies. The data was processed and cleaned for analysis and then divided into training and testing datasets to build and evaluate the predictive model. The machine learning algorithms and statistical methods used to build the model include decision trees, random forests, and logistic regression.

The results of this study will provide the insurance company with a valuable tool that can be used to optimize its revenue streams and improve its business model through targeted marketing and sales efforts. The findings of this research will also have implications for other industries that are looking to improve their sales and revenue through predictive modeling. Additionally, the results of this study will contribute to the larger body of literature on customer behavior and decision-making in the insurance industry. Less than half of our country's people have health insurance and 60% of people who own their vehicles don't get insurance [1]. The graphs below show the growth in the number of people covered by health insurance and the growing sales of automobiles. This opens a vast potential market for insurance companies as many more people might just be interested in buying insurance for their vehicles as well. This presents itself as a huge opportunity for an insurance company that wants to sell people both health and vehicle insurance. The statistics show that a tiny chunk of people hold both health and vehicle insurance. Therefore, using a model to analyze the customer data and trends to predict customer behavior in buying auto insurance for their vehicle will prove highly beneficial for the company.

II. Literature Review

The insurance industry has been undergoing significant changes in recent years, with technological advancements and increasing customer demands leading to a shift toward data-driven decision-making. In particular, data science techniques, such as machine learning, have emerged as valuable tools for predicting customer interests and preferences. Several studies have explored the use of data science techniques in the insurance industry.

For instance, a study by Bhatia and Singh (2017)[2] used data mining techniques to predict customer churn in the Indian insurance market. The study found that decision tree algorithms, such as Random Forests, were the best performing in terms of accuracy and precision.

A study by Abbasi et al. (2019)[3] used data mining and machine learning techniques to predict customer preferences in the insurance industry.The study employed three different algorithms, including Decision Trees, Random Forests, and Support Vector Machines (SVMs), and found that SVMs performed the best. More recently, studies have focused on using machine learning techniques to predict customer interests in specific insurance products, such as vehicle insurance.

A study by Gao et al. (2019)[4] used machine learning algorithms to predict customer interest in buying car insurance. The study found that logistic regression, decision tree, and random forest models were effective in predicting customer behavior. In conclusion, the literature suggests that data science techniques, particularly machine learning, can be valuable tools for predicting customer interests and preferences in the insurance industry. Several studies have demonstrated the effectiveness of decision tree algorithms, such as Random Forests, in predicting customer behavior. Furthermore, recent studies have highlighted the potential of machine learning techniques for predicting customer interests in specific insurance products, such as vehicle insurance. The current research builds on this literature by using data science techniques to predict customer interest in purchasing vehicle insurance from a health insurance company.

III. Materials And Methods

The process of predicting required information from certain data consists of 4 major stages:

STAGE 1

Data Preprocessing - processing and cleaning of the health insurance holders' data.

STAGE 2

Data Visualization and Representation of the processed data for identifying features and attributes as well as extracting more information from the visualizations.

STAGE 3

Classification involves the application of different machine-learning algorithms to build models that will make the required predictions.

STAGE 4

Evaluation of trained model based on different performance metrics. This tells us the efficiency and effectiveness of the model.

A. Dataset

The dataset used for this research paper contains information on more than 38000 people who own health insurance in a company. The trained model would predict whether these people would be interested in buying vehicle insurance from the same company. There are several features and attributes including Gender, Age, Driving License, and Region Code, if the customer previously, Vehicle Age, Vehicle Damage, Annual

Premium, Policy Sales Channel, Vintage, and the Response of the customers in the training dataset upon which the algorithm - Logistic Regression, Random Forest, and Naive Bayes Algorithm will be applied.

Table I: Descriptive Analysis of the Dataset

	Age	Driving License	Annual Premium	Policy Sales Channel	Vintage
Count	381109	381109	381109	381109	381109
Mean	38.82	0.997	30564.38	112.03	154.347
Standard Deviation	15.51	0.046	17213.15	54.20	83.67
Minimum	20	0	2630	1	10
Maximum	85	1	540165	163	299

B. Data Analysis

After evaluating individual columns of the dataset as shown in Table I, a few steps are taken to check whether the data requires cleaning, this may involve removing any null values or dropping any unnecessary columns.

```
RangeIndex: 381109 entries, 0 to 381108
Data columns (total 12 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   id                    381109 non-null  int64
 1   Gender                381109 non-null  int64
 2   Age                   381109 non-null  int64
 3   Driving_License       381109 non-null  int64
 4   Region_Code           381109 non-null  float64
 5   Previously_Insured    381109 non-null  int64
 6   Vehicle_Age           381109 non-null  int64
 7   Vehicle_Damage        381109 non-null  int64
 8   Annual_Premium        381109 non-null  float64
 9   Policy_Sales_Channel  381109 non-null  float64
 10  Vintage               381109 non-null  int64
 11  Response              381109 non-null  int64
dtypes: float64(3), int64(9)
```

Fig. 1. Data Information

Figure 1 shows information about the different attributes of the dataset and if there are any null values present. Since there are no null values, data is checked for duplicates as well.

```
[15] df.duplicated().sum()

0
```

Fig. 2. Checking for Duplicates

Further, a heatmap is created to get the correlations of different columns with each other and identify patterns if any as shown in Figure 3.

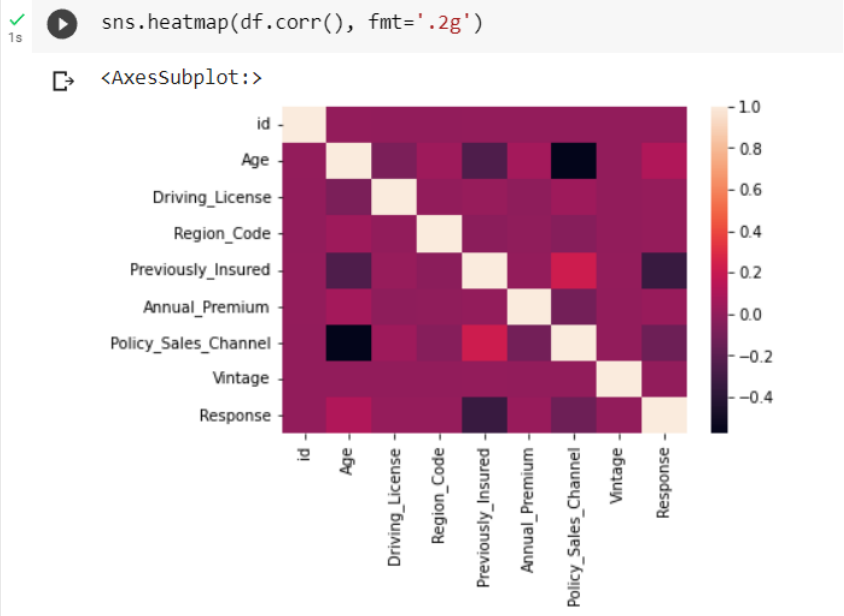


Fig. 3. Heatmap

Figure 4 shows the correlation of different attributes with respect to the response variable. While Vintage, Policy Sales Channel, Previously Insured, and Gender are negatively correlated to the Response attribute, the remaining attributes are positively correlated.

```
Out[50]:
Gender                -0.052440
Age                    0.111147
Driving_License        0.010155
Region_Code            0.010570
Previously_Insured    -0.341170
Vehicle_Age            0.221874
Vehicle_Damage         0.354400
Annual_Premium         0.022575
Policy_Sales_Channel  -0.139042
Vintage               -0.001050
Response               1.000000
Name: Response, dtype: float64
```

Fig. 4. Correlation of attributes with respect to outcome

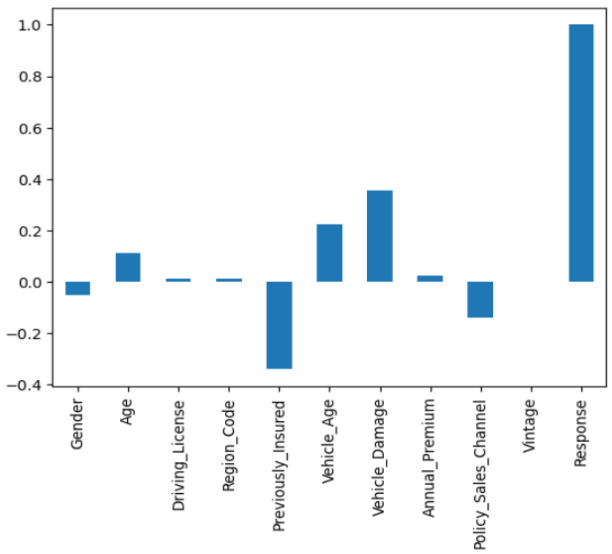


Fig. 5. Plot of correlation of attribute with respect to outcome

Sometimes when dealing with non-numeric data, it becomes important to convert it into numeric data so that it can be used in machine learning models.

]:

	Gender	Vehicle_Age	Vehicle_Damage
count	381109	381109	381109
unique	2	3	2
top	Male	1-2 Year	Yes
freq	206089	200316	192413

Fig. 6. Nonnumeric columns in data

Gender, Vehicle Age, and Vehicle Damage all are important attributes for predicting customer interest but need to be converted into numeric values.

```
[17] df['Vehicle_Age'].replace(['> 2 Years', '1-2 Year', '< 1 Year'],
                               [2, 1, 0], inplace=True)

[18] df['Vehicle_Damage'].replace(['Yes', 'No'],
                               [1, 0], inplace=True)

[19] df['Gender'].replace(['Male', 'Female'],
                               [1, 0], inplace=True)
```

Fig. 7. Converting nonnumeric to numeric columns

```
df.Gender.value_counts().plot(kind='bar')
```

```
<AxesSubplot:>
```

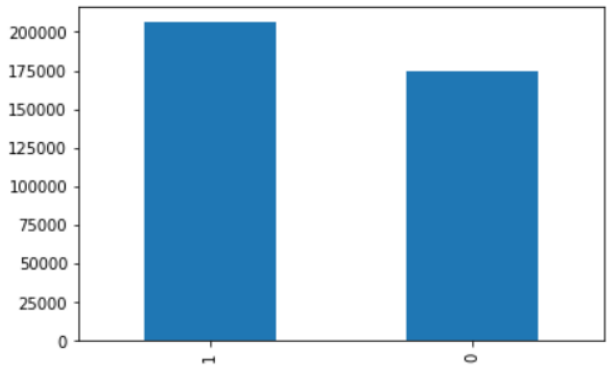


Fig. 8. Count plot of Gender, 1= Male and 0 = Female

On plotting the plot for whether a person has a driving license or not, it was found that 380297 people had driving licenses while the remaining 812 didn't have one.

```
]:
1    380297
0       812
Name: Driving_License, dtype: int64
```

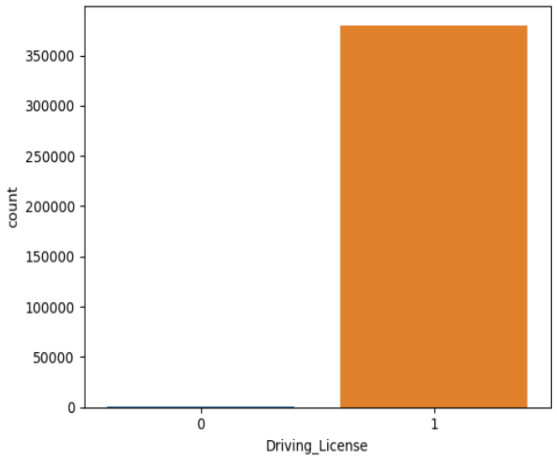


Fig. 9. Count plot for Driving License, 1 = people with driving license, 0 = people without a license

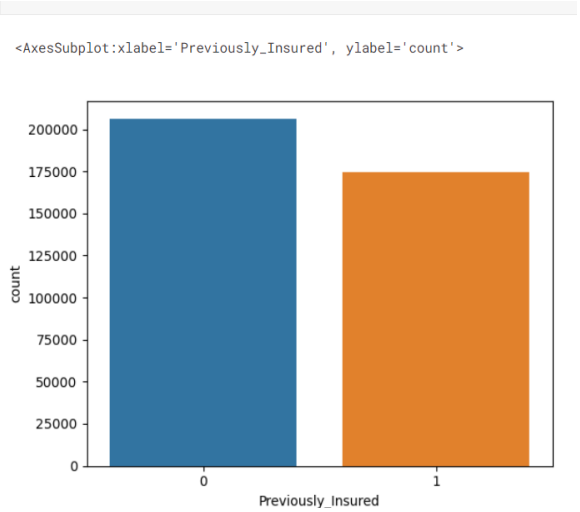


Fig .10. Count plot of previously insured

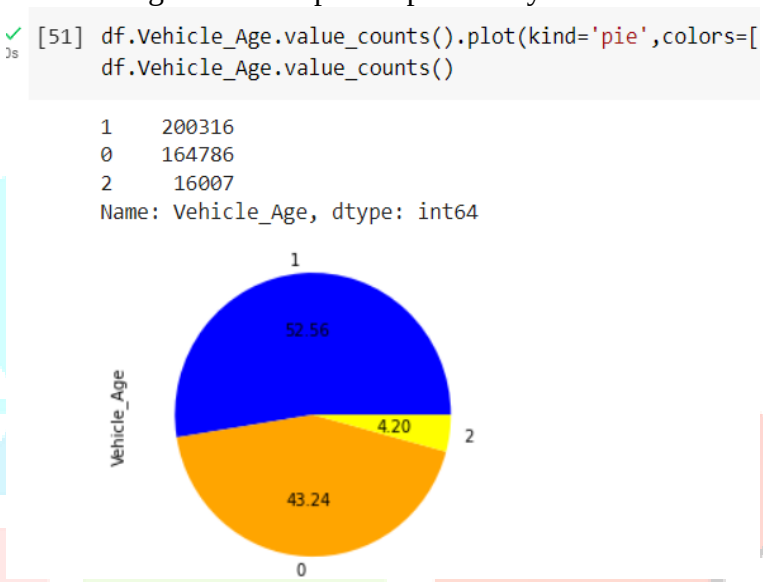


Fig. 11. Pie chart of vehicles owned.

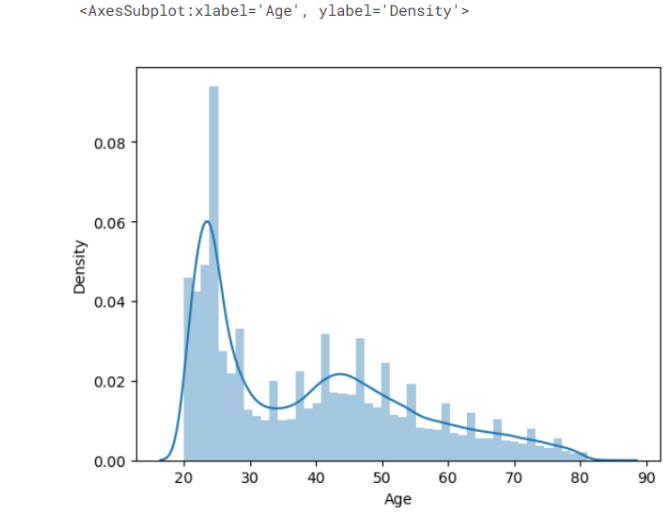


Fig. 12. Distribution of health insurance bought by people of different ages.

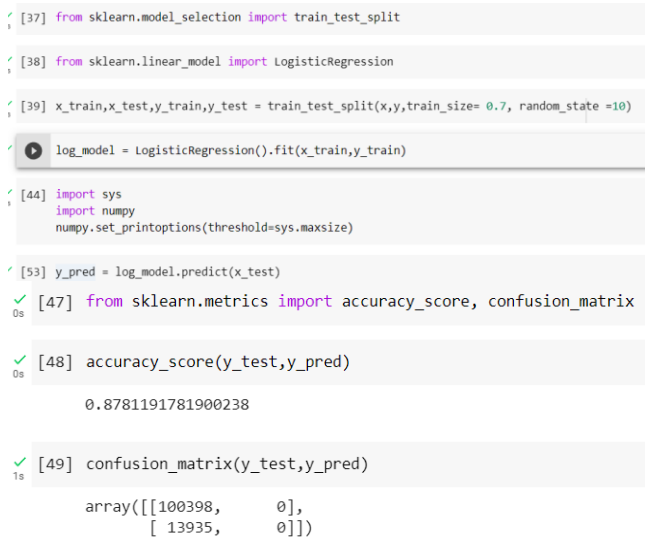
IV. Model Training

Now that the dataset is processed and visualized to draw out the necessary information from it, machine learning models are trained using it so that predictions can be made.

The data is divided into train and test sets, the train set is used to train our model while the test set is used to evaluate the trained model. This is done using different performance measures and metrics.

A. Logistic Regression

Applying Logistic Regression on the data using the sci-kit learn [5], a library designed for data science in Python. Logistic regression classifies the data and gives output in the form of a sigmoid curve. The values above and below a particular threshold are decided as the two classes the data would be classified into.



```
[37] from sklearn.model_selection import train_test_split

[38] from sklearn.linear_model import LogisticRegression

[39] x_train,x_test,y_train,y_test = train_test_split(x,y,train_size= 0.7, random_state =10)

     log_model = LogisticRegression().fit(x_train,y_train)

[44] import sys
     import numpy
     numpy.set_printoptions(threshold=sys.maxsize)

[53] y_pred = log_model.predict(x_test)

[47] from sklearn.metrics import accuracy_score, confusion_matrix

[48] accuracy_score(y_test,y_pred)

     0.8781191781900238

[49] confusion_matrix(y_test,y_pred)

     array([[100398,      0],
            [ 13935,      0]])
```

Fig. 13. Application of Logistic Regression

B. Naïve Bayes

The Naive Bayes algorithm is applied to the data next. Naive Bayes is a probabilistic classification algorithm that is widely used in machine learning for solving classification problems. It is based on the Bayes theorem, which states that the probability of a hypothesis (or label) given some evidence (or features) is proportional to the probability of that evidence given the hypothesis, multiplied by the prior probability of the hypothesis.

```
[58] from sklearn.naive_bayes import GaussianNB
     model = GaussianNB()
     model.fit(x_train,y_train)

     GaussianNB()

[61] y_pred = model.predict(x_test)
     y_pred

[62] print('Accuracy:%d',(model.score(x_test,y_test)))

     Accuracy:%d 0.8204367942763682

[63] from sklearn.metrics import confusion_matrix
     confusion_matrix = confusion_matrix(y_test,y_pred)
     print(confusion_matrix)

     [[88744 11654]
      [ 8876  5059]]
```

Fig. 14. Application of Naïve Bayes algorithm

C. Random Forest

Applying the Random Forrest algorithm on the data, this is a type of ensemble learning algorithm(bagging) that divides the data into subsets and then applies decision trees on the subsets individually. The mean of the result from these Decision trees is then taken and evaluated.

```
[54] from sklearn.ensemble import RandomForestClassifier
     clf = RandomForestClassifier()
     clf.fit(x_train,y_train)

     RandomForestClassifier
     RandomForestClassifier()

[59] y_pred = clf.predict(x_test)
     y_pred

[56] print('Accuracy:%d',(clf.score(x_test,y_test)))

     Accuracy:%d 0.8679296441097496

[57] from sklearn.metrics import confusion_matrix
     confusion_matrix = confusion_matrix(y_test,y_pred)
     print(confusion_matrix)

     [[97547  2851]
      [12249  1686]]
```

Fig. 15. Application of random forest algorithm

V. MODEL EVALUATION

The applied machine learning models have performance measures and metrics that are used to evaluate a particular model's performance. A confusion matrix provides a summary of the predictions made by the model on a set of test data, compared to the true class labels.

Table II: Confusion Matrix

	Predicted Positive	**Predicted Negative**
True Positive	TP	FN
True Negative	FP	TN

Using the confusion matrix, the Accuracy, Precision, Recall, and F1 score of a machine learning model is calculated. These measures are used to evaluate an applied model and the validity of the results it gives.

A. Evaluating Logistic Regression

Performance Metric is evaluated as follows:

```
#Logistic Regression
print("Accuracy Score: ",accuracy_score(y_test,y_pred))
print("Precision Score: ", precision_score(y_test, y_pred,average='macro'))
print("Recall Score: ", recall_score(y_test, y_pred,average='macro'))
print("F1 Score: ", f1_score(y_test, y_pred,average='macro'))

Accuracy Score:  0.8781191781900238
Precision Score:  0.4390595890950119
Recall Score:  0.5
F1 Score:  0.4675524260586501
```

Fig. 16. Logistic Regression Performance metric

B. Evaluating Naïve Bayes

Table III: Confusion Matrix for Naïve Bayes

	Predicted Positive	**Predicted Negative**
True Positive	88744	11654
True Negative	8876	5059

```
#Naive Bayes
print('Accuracy:%d',(model.score(x_test,y_test)))
print("Precision Score: ", precision_score(y_test, y_pred,average='macro'))
print("Recall Score: ", recall_score(y_test, y_pred,average='macro'))
print("F1 Score: ", f1_score(y_test, y_pred,average='macro'))

Accuracy:%d 0.8204367942763682
Precision Score:  0.4390595890950119
Recall Score:  0.5
F1 Score:  0.4675524260586501
```

Fig. 17. Naïve Bayes Performance metric

C. Evaluating Naïve Bayes

Table IV: Confusion Matrix for Random Forest

	Predicted Positive	**Predicted Negative**
True Positive	97547	2851
True Negative	12249	1686

```
#Random Forests
print('Accuracy:%d',(clf.score(x_test,y_test)))
print("Precision Score: ", precision_score(y_test, y_pred,average='macro'))
print("Recall Score: ", recall_score(y_test, y_pred,average='macro'))
print("F1 Score: ", f1_score(y_test, y_pred,average='macro'))

Accuracy:%d 0.8673698757139233
Precision Score:  0.6288798411833533
Recall Score:  0.5473993666670591
F1 Score:  0.5568800823740427
```

Fig. 18. Random Forest Performance metric

VI. Comparison Of Models

On Comparing the Metrics of the three applied models, it is found that while Random Forests and Logistic Regression both perform similarly in accuracy, Random Forests have better precision and recall as well as a higher F1 score implying a better prediction of both true positives and negatives while minimizing the false positives and negatives.

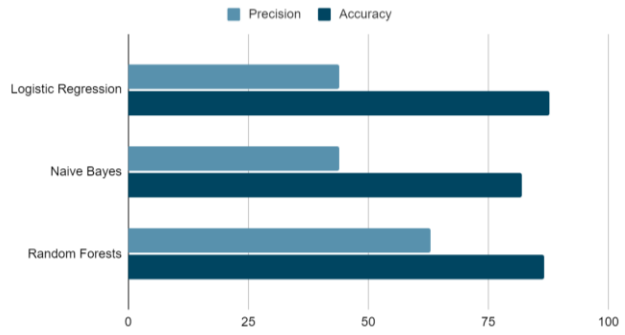


Fig. 19. Comparing Performance Metrics (Precision Vs Accuracy)

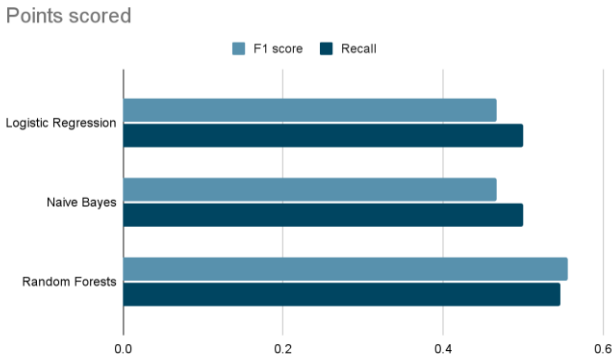


Fig. 20. Comparing Performance Metrics (F1 score Vs Recall)

Naive Bayes on the other hand performs worse overall when compared to the other two models. Although Logistic Regression has greater accuracy, it's a minimal increment over Random Forests, and since Random Forests have better overall performance, it predicts customer interests better. Figure 17 shows the performance metrics of all the applied models.

VII. Conclusion

In conclusion, this research paper demonstrates the effective use of data science techniques for predicting customer interest in purchasing vehicle insurance from a health insurance company. By applying data processing, analysis, and visualization techniques, we were able to prepare the data for machine learning models, which were trained and evaluated for their predictive performance. The results suggest that the random forests model is the best-performing model, as it achieved the best accuracy, precision, and F1 score.

The study contributes to the growing field of data science in the insurance industry and highlights the potential benefits of using machine learning models to predict customer interests which can provide valuable insights into customer interests and preferences, helping companies to design more effective marketing strategies and improve customer satisfaction.

Furthermore, the research builds on the existing literature by demonstrating the effectiveness of machine learning techniques in predicting customer interests in specific insurance products, such as vehicle insurance. The results of this research can aid insurance companies in developing targeted marketing strategies and improving customer satisfaction. The future scope of this research is to extend beyond a single company to compare customer interests in vehicle insurance across multiple companies.

Additionally, exploring the factors that influence customer interest in purchasing insurance products can provide a better understanding of customer behavior and help companies design more effective marketing strategies. The application of advanced machine learning techniques, such as deep learning, can also improve the accuracy and precision of the predictive models, providing more reliable insights into customer interests. Finally, incorporating customer feedback and opinions can provide a more comprehensive view of customer preferences and help companies tailor their services accordingly.

References

[1] Around 57% of vehicles in India uninsured - Times of India, available at - https://timesofindia.indiatimes.com/business/india-business/57-vehicles-in-india-uninsured/articleshow/79652274.cms

[2] Bhatia, V., & Singh, H. (2017). Predicting customer churn in Indian insurance industry using data mining techniques. Journal of Advances in Management Research, 14(2), 217-238. https://doi.org/10.1108/JAMR-07-2016-0061

[3] Abbasi, A., Khan, A., & Ahmad, R. (2019). Customer preferences prediction in the insurance industry using data mining and machine learning techniques. International Journal of Computer Science and Information Security, 17(2), 1-6. http://dx.doi.org/10.5281/zenodo.2545988

[4] Gao, C., Zhang, H., & Wang, L. (2019). Predicting the willingness of purchasing car insurance based on customer behavior. Advances in Intelligent Systems and Computing, 921, 309-320. https://doi.org/10.1007/978-981-13-3115-2_30