PREDICTION OF VEHICLE INSURANCE CLAIMS USING MACHINE LEARNING ALGORITHMS

CIS4035-N-FJ1-2021

MACHINE LEARNING

SCHOOL OF COMPUTING, ENGINEERING AND DIGITAL TECHNOLOGIES

MAY 20, 2022

Prepared by

B1326237 - GOWTHAMI NAGAPPAN

Gowthami Nagappan
Teesside University - B1326237@live.tees.ac.uk

TABLE OF CONTENTS

Gowthami Nagappan
Teesside University - B1326237@live.tees.ac.uk

ABSTRACT

The increasing number of vehicle insurance claims required the development of strategic approaches for managing individual claims. One of efficient way for fixing such situation is using machine learning (ML) Techniques. To improve the quality-of-service and timely delivery, vehicle insurers have started to incorporate and apply machine learning techniques to improve the perception and ability of the data for efficiency. so the requirements of the customers can be met efficiently. This research work explains on how the insurance firms employ machine learning into their operations and how these models may be applied for massive amount of data in the insurance industry. To forecast the claims, we use ML Regression methods such as Multiple linear regression, Decision Trees, Random Forest and SVM to predict ultimate insurance claims. The performance of different models is evaluated and compared with one another. The results show that RF (Random Forest Regression) is better when compared to other regression methods with a good accuracy and less errors respectively.

Gowthami Nagappan
Teesside University - B1326237@live.tees.ac.uk

INTRODUCTION

Background

Every year, the average expenses of vehicle insurance keeps rising. The insurance firms are focused on analyzing insurance data to anticipate future insurance claims. When a policyholder (customer) makes a request to the insurance company for the policy coverage or compensation, it is known as a vehicle insurance claim. The concerned firms should check the request before deciding whether to pay their customer or policyholder. Automobile insurance quotes are determined by several factors. In particular, accurate claim cost prediction is critical in establishing policy premium.

So, the firms can avoid potential customer loss due to overcharging and potential profit loss owing to undercharging. Our primary goal is to develop a ML model which will accurately determine the level of claim. As a result, the model should examine the consumers information in the dataset such as the type of claim, claim number, reason for claim, first claim payment and other critical factors. The results of the model will show how much insurance companies will pay out on total claims.

REVIEW OF RELEVANT LITERATURE

Prediction of vehicle insurance claims using various technologies

Machine learning algorithms are employed in analyzing the insurance market because the insurance firms have a real desire to automate their operations. [1] Wijegunasekara and Weerasinghe studied various machine learning algorithms to forecast claims in 2016 and other researchers have addressed the problem of prediction in insurance sector using ML models. [2] The neural networks were shown to be the most accurate predictor. The thesis "Research on probability-based Learning application on car insurance data" is another example of a similar and successful issue (Jing et al. 2018). [3] Bayesian network was used to determine whether or not a claim has been submitted, whereas Kowshalya and Nandhini (2018) used machine learning approaches to forecast fake claims and compute the claims depending on personal information.

[4] A model that forecast the intensity of the claim as well as the expenses of fixing vehicle damage is yet another example of insurance sector analytics (Dewi et al. 2019). [5] This shows how the firms delve in-depth on methods of applying ML to their client's data. Singh et al. 2019 demonstrated a process that uses crashed car images as inputs. It generates exact data such as repair costs to estimate the cost of an insurance claim. [6] Rather than predicting the occurrence of insurance claims, this research focuses on calculating repair costs. (Stucki 2019).

Gowthami Nagappan
Teesside University - B1326237@live.tees.ac.uk

Furthermore, [7] a model was developed for predicting the claims (Abdelhadi et al. 2020); We can say that some of the prior research in the insurance sector used [8] machine learning models which showed that XGBoost model is better for risk evaluation in the industry (Pesantez-Narvaez et al. 2019; Abdelhadi et al. 2020).

In our research, we analyzed insurance claim data with 50000 attributes and found that Random Forest performs much better than decision trees. Our findings demonstrate that, all the regression models utilized in this study, SVM is the weakest at predicting claim occurrence.

EXPERIMENT

Four ML regression methodologies are used to predict the insurance claims. Linear Regression, Decision Tree, Random Forest Regression and (SVM) Support Vector Machine predictions were utilized for creating a predictive model since they fit well with how python programming tools are used to formulate problems.

The Dataset

The Dataset was taken from Kaggle (https://www.kaggle.com/competitions/actuarial-loss-estimation/data?select=test.csv) which had 90,000 realistic, synthetically generated worker compensation insurance policies, all of which have been involved in an accident. There is a demographic & worker information as well as a text description of the accident for each report.

Train.csv - The training set containing 54,000 insurance policies were used to train the model.
Test.csv - The Test dataset to predict the total payment of insurance claim as a result.
Sample_submission.csv - A sample file in the proper format for submission.

```
Train_Dataset.head()
```

	ClaimNumber	DateTimeOfAccident	DateReported	Age	Gender	MaritalStatus	DependentChildren	DependentsOther	WeeklyWages	PartTimeFullTime	HoursWorkedPerWeek	DaysWorkedPerWeek	ClaimDescription
0	WC8285054	2002-04-09T07:00:00Z	2002-07-05T00:00:00Z	48	M	M	0	0	500.00	F	38.0	5	LIFTING TYRE INJURY TO RIGHT ARM AND WRIST INJURY
1	WC6982224	1999-01-07T11:00:00Z	1999-01-20T00:00:00Z	43	F	M	0	0	509.34	F	37.5	5	STEPPED AROUND CRATES AND TRUCK TRAY FRACTURE ...
2	WC5481426	1996-03-25T00:00:00Z	1996-04-14T00:00:00Z	30	M	U	0	0	709.10	F	38.0	5	CUT ON SHARP EDGE CUT LEFT THUMB
3	WC9775968	2005-06-22T13:00:00Z	2005-07-22T00:00:00Z	41	M	S	0	0	555.46	F	38.0	5	DIGGING LOWER BACK LOWER BACK STRAIN
4	WC2634037	1990-08-29T08:00:00Z	1990-09-27T00:00:00Z	36	M	M	0	0	377.10	F	38.0	5	REACHING ABOVE SHOULDER LEVEL ACUTE MUSCLE STR...

Fig : Overview of Dataset

Gowthami Nagappan
Teesside University - B1326237@live.tees.ac.uk

Data Preprocessing

The Data preprocessing was done by renaming the columns and changing the datatype of each column. Then the missing values and null values were checked in the Dataset. Because some of the columns had missing data, the gaps were filled using mean and mode imputation.

Data Transformation

Transforming the Age column into ‘young’, ‘Middle-Age’ and ‘Old’ based on Min and Max factors. Also categorized the weekly wages into ‘Low', 'Below Average', 'Average Wage', 'Above Average' and 'High' to understand better in data analysis.

EXPLORATORY DATA ANALYSIS

1. Univariate Analysis

In Univariate Analysis, the below plot shows that the data for Target Variables of Ultimate Incurred Claim Cost and Initial Incurred Claim Cost are right skewed.

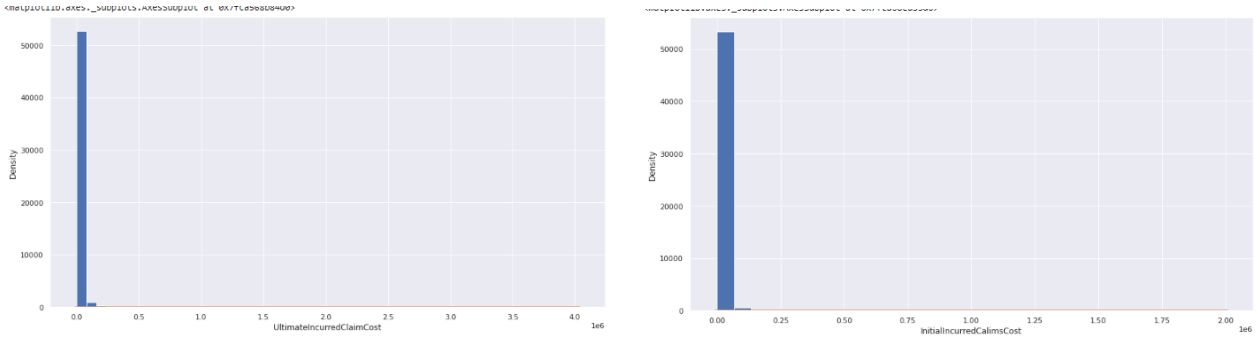

Fig : Target Feature Skewness

The below 2 plots shows that, the claims are higher from the middle age group (25-40) which is closer to 57% when compare with young and old age.

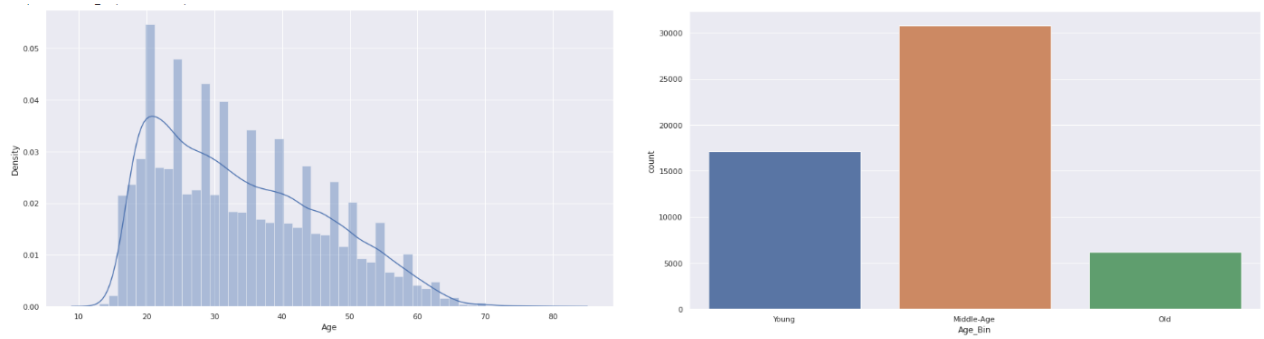

Fig : Age Based Analysis

Gowthami Nagappan
Teesside University - B1326237@live.tees.ac.uk

In the below plot, it's evident that the claims made from people with no children as dependents are high (close to 94%). The claims made from people with no other dependents is higher than people with other dependents (more than 99%).

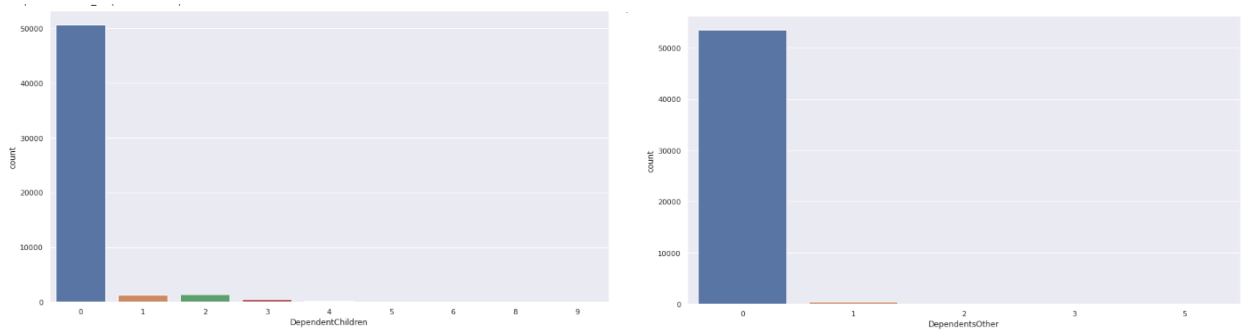

Fig : Dependents Based Analysis

The Below images shows that, the data for hours worked per week is positively skewed and other image states that high number of claims made by people who worked for more than 5 days per week (91%).

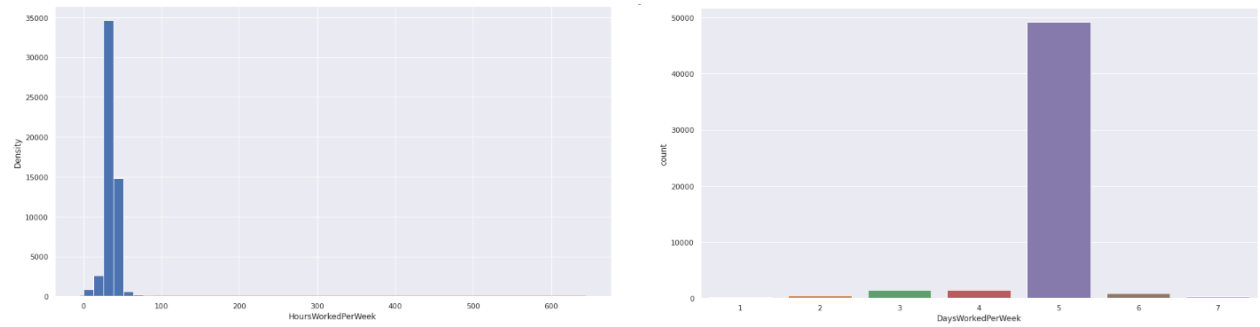

Fig : Analysis based on Hours of Work

From the below graph, we can see the categorical columns of gender, marital status and full/part time job columns. There are more no of male (77%) who claimed high insurance compared with females (22%). Almost 48% people who are single claimed more insurance compared with married and unmarried. Even the people who hold full time jobs, their claim were higher at 91% .

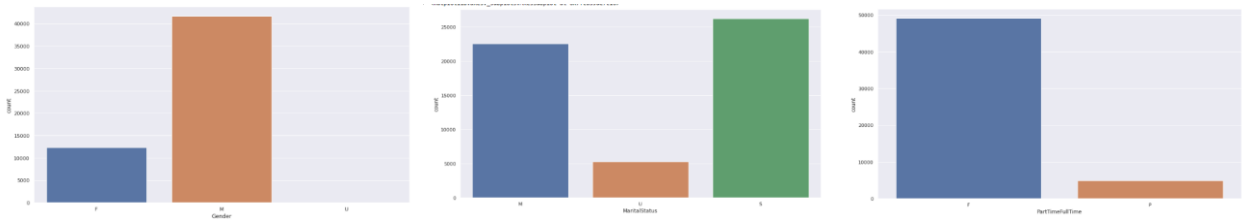

Fig : Analysis based on Categorical Columns

Gowthami Nagappan
Teesside University - B1326237@live.tees.ac.uk

2. Bivariate Analysis

In Bivariate analysis, two features of the dataset were taken into analysis. The old age group people (50-80) got high payment claims from the insurance company. In addition, the people who have more number of children as dependents also got high claim payments from the insurance company.

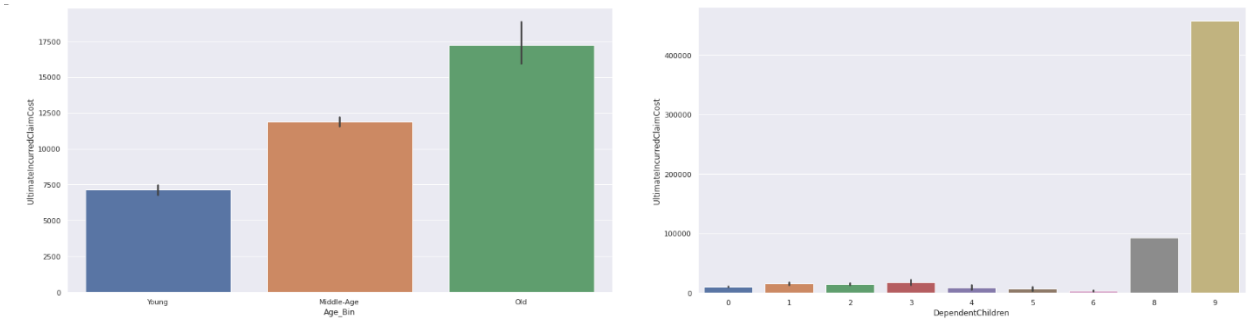

Fig : Analysis with Age , Dependents Vs Target Variable

The below images states that, the high payment claimed by people are those with the wages are above average.

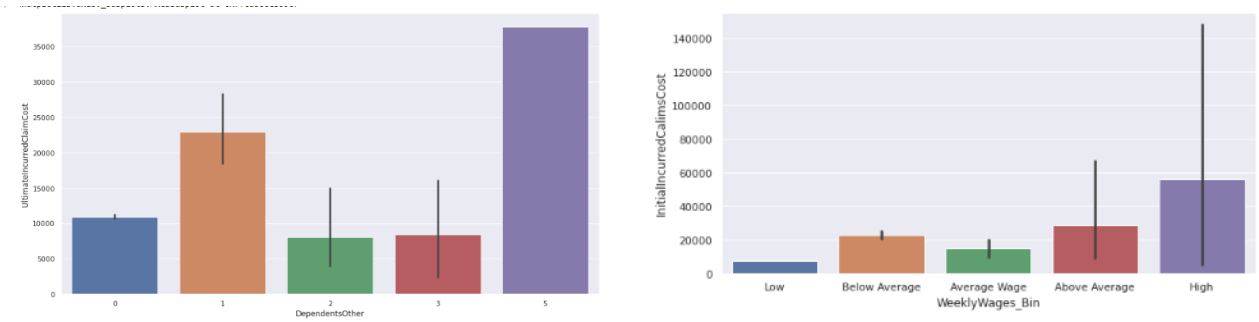

Fig: Analysis with Other Dependents and Wages Vs Target Variable

The below Graph states that, People whose wages are below average and above average got high claims payments by the insurance company.

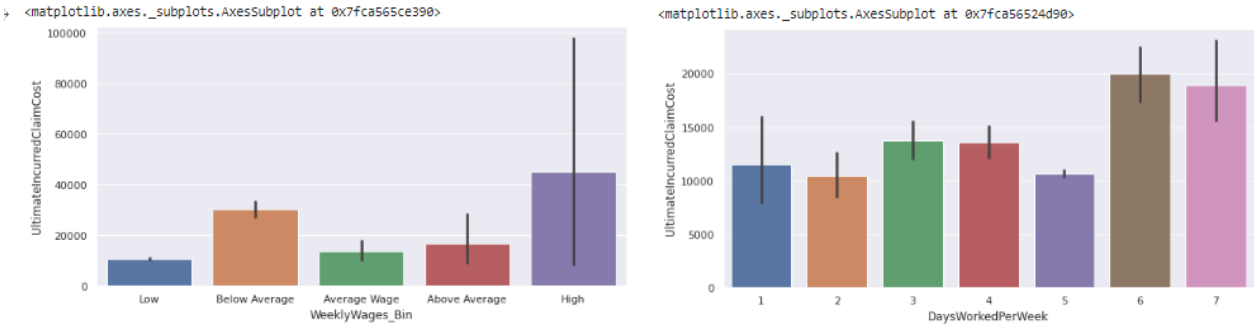

Fig: Ages , Days Worked Vs Target Variable

Gowthami Nagappan
Teesside University - B1326237@live.tees.ac.uk

3. Outlier Analysis

In outlier analysis, the below boxplot shows that there are a lot of outliers in "Initial Incurred Claims Cost" and "Ultimate Incurred Claim Cost".

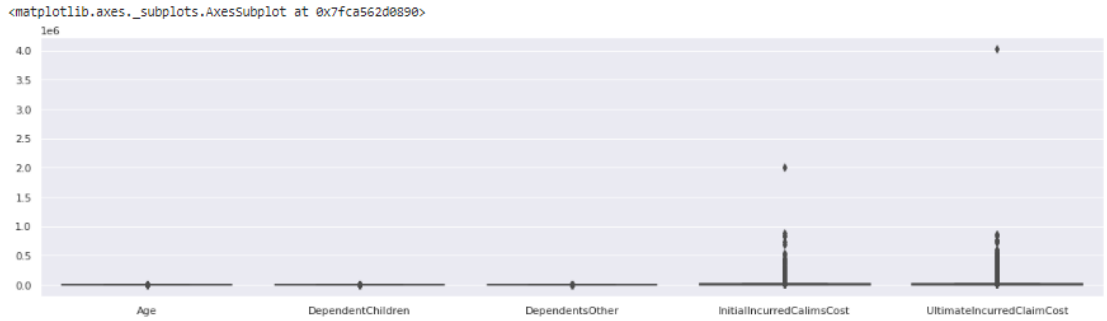

Fig : Outlier Analysis

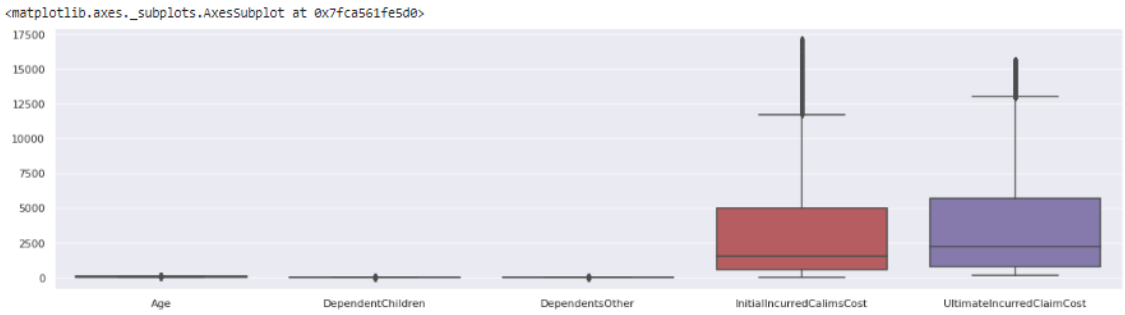

Fig : Outlier Analysis

The outliers present in the target feature has dropped and normalized using functions to avoid incorrect measured data.

RESEARCH QUESTION

The goal of this study is to examine total insurance claim payments and develop a machine learning model which can accurately forecast the claim payment. Regression analysis was used as the dataset contains both dependent and independent variables. There is also a date and time column since the Insurance claim volume is reported on a yearly basis.

FEATURE SELECTION

To improve the learning capacity and quality of the model, the non-correlated input features are reduced so the models don't get confused. Also, we have removed some of the features in the dataset such as Claim Number, Date of the Accident, Date Reported, Marital Status and Claim Description Columns. After reducing noise from the dataset, we can split the data into train and test to evaluate.

Gowthami Nagappan
Teesside University - B1326237@live.tees.ac.uk

Data preparation

The Target Variable "Ultimate Incurred Claim Cost" Column was dropped from the Train dataset to split Training and Validation Data. 70% of the data was taken for training and 30 % of the data for test validation (From Train Dataset).
New Unseen data has been used for Final Claim prediction by the model which has good accuracy (Test Dataset).

METHODOLOGY

1. Linear Regression

The multiple linear connections between a response variable (target) and a collection of predictor variables are estimated using the linear regression method. When the target variable is binary, linear regression is not acceptable (Sabbeh 2018). In linear regression, best fit line is generated using mean squared error which determines the relationship between the dependent, Y, and independent, X.

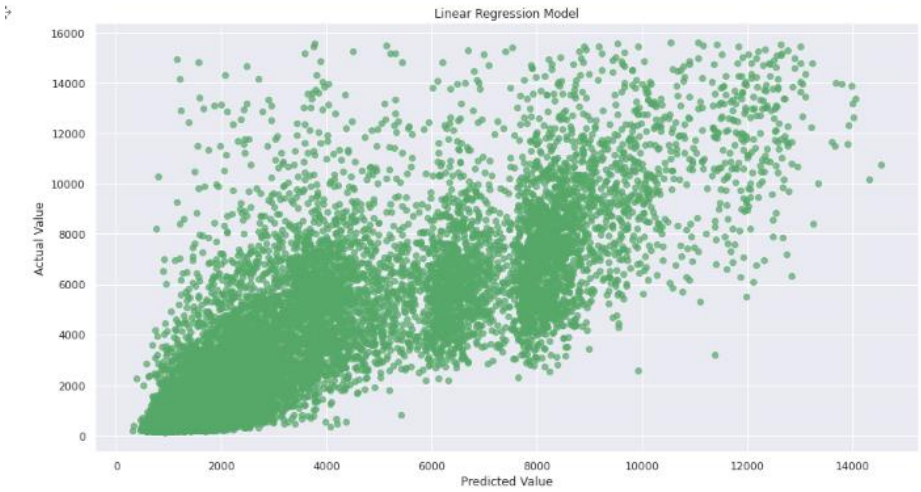

Fig : Linear Regression

For predicting associations between continuous variables, linear regression is a simple and easy statistical regression procedure. Linear regression illustrates a linear relationship between an independent and dependent variable (X-axis & Y-axis). As there are various input variables, it is called multiple linear regressions. A slanted straight line is constructed by the linear regression model to represent the relationship between the variables.

Gowthami Nagappan
Teesside University - B1326237@live.tees.ac.uk

```
LinearRegression()
```

```
#Checking the train score
firstmodel.score(x_train,y_train)
print("Training set accuracy: ", +(firstmodel.score(x_train, y_train)))

#Checking the test score
firstmodel.score(x_test,y_test)
print("Test set accuracy:" , +(firstmodel.score(x_test, y_test)))
```

```
Training set accuracy:  0.6798255481178431
Test set accuracy: 0.6854008835218999
```

Fig: Linear Model Training and Testing Accuracy

The model was created and trained using train dataset. The train and test accuracy displayed in above image has Training accuracy of 67% and Validation at 68%.

```
#Predictions on the test data set.
y_pred = firstmodel.predict(x_test)

print('The predict values are:\n',y_pred)
```

```
The predict values are:
 [ 3191.13789665 10054.30742874  1981.2219766  ...  4914.80242607
  1362.63235906  1414.83045855]
```

Fig : The prediction of Target Values

2. Decision Tree Regression

The root and leaf nodes of the decision tree are represented by intermediate nodes, which represent non-leaf nodes. In decision-making, the root node is picked as the data item having the highest priority. The decision tree's split procedure is governed by the respective nodes.

During the training phase, the decision tree learns and its effectiveness is assessed during the testing phase. The breadth and depth of DT's testing and training data has a dynamic impact on the prediction's performance and efficiency.

Gowthami Nagappan
Teesside University - B1326237@live.tees.ac.uk

```
DecisionTreeRegressor()
```

```
#Checking the train score
secondmodel.score(x_train,y_train)
print("Training set accuracy: ", +(secondmodel.score(x_train, y_train)))

#Checking the test score
secondmodel.score(x_test,y_test)
print("Test set accuracy:" , +(secondmodel.score(x_test, y_test)))
```

```
Training set accuracy:  1.0
Test set accuracy: 0.516210769351955
```

Fig: Decision Tree Training and Testing Accuracy

The train and test accuracy are displayed in the above image where the training accuracy was 100% and the Validation has 51%. The below image displays the predicted results of insurance claim.

```
# predicting the test set results
y_pred1= secondmodel.predict(x_test)
print('The predict values are:\n',y_pred1)
```

```
The predict values are:
 [3218.958918  8579.471902  2115.35383   ... 4848.11682    452.1326662
  199.3743564]
```

Fig : The prediction of Target Values

3. Random Forest Regression

Random forest is an algorithm based on tree model which form trees using bagging. These are coached on their own. The prediction is the result of a series of orthogonal split decisions in the multivariate space of variables. Each decision tree is built using samples chosen at random. When a feature is used as a part in a random decision tree, that is no longer the best of all features; instead, the best bootstrap factor is constructed at random.

```
#Checking the train score
thirdmodel.score(x_train,y_train)
print("Training set accuracy: ", +(thirdmodel.score(x_train, y_train)))

#Checking the test score
thirdmodel.score(x_test,y_test)
print("Test set accuracy:" , +(thirdmodel.score(x_test, y_test)))
```

```
Training set accuracy:  0.9509445717296763
Test set accuracy: 0.726165582100248
```

Fig: Random Forest Tree Training and Testing Accuracy

Gowthami Nagappan
Teesside University - B1326237@live.tees.ac.uk

The model was created and trained using train dataset. The train and test accuracy displayed in above image has a Training accuracy of 95% and Validation is 72%.The below image displays the Predicted result of insurance claim.

```
# predicting the test set results
y_pred2= thirdmodel.predict(x_test)
print('The predict values are:\n',y_pred2)
```

```
The predict values are:
 [3258.3329582  9237.0789959  1529.65064538 ... 7384.3224032   665.62826035
  381.56254478]
```

Fig : The prediction of Target Values

4. Support Vector Machine (SVM) Regression

SVM's is a sort of frontier hyperplane which separates the class by finding the biggest margin between the closest points in a training set of any attribute. The kernel method allowed elements in a limited-dimensional space to be projected on a larger space, allowing them to be segregated linearly regardless of the dimensional space [10]. SVM has been proven to be highly accurate in a range of applications, including cancer diagnosis. [11].

```
SVR()
```

```
#Checking the train score
fourthmodel.score(x_train,y_train)
print("Training set accuracy: ", +(fourthmodel.score(x_train, y_train)))

#Checking the test score
fourthmodel.score(x_test,y_test)
print("Test set accuracy:" , +(fourthmodel.score(x_test, y_test)))
```

```
Training set accuracy:  0.1949175403714759
Test set accuracy: 0.19737413491140432
```

Fig: SVM Training and Testing Accuracy

The model was created and trained using train dataset. The train and test accuracy are displayed in the above image where the Training accuracy is very low at 19 % and Validation at 19%.The below image displays the predicted result of insurance claim.

```
# predicting the test set results
y_pred3 = fourthmodel.predict(x_test)
print('The predict values are:\n',y_pred3)
```

```
The predict values are:
 [2452.52791317 4075.1819088  2213.84214469 ... 2859.12818538 1812.74217266
 1989.50855096]
```

Fig : The prediction of Target Values

Gowthami Nagappan
Teesside University - B1326237@live.tees.ac.uk

MODEL EVALUATION METRICS

The result of each regression model is predicted by using various regression metrics like R2, MAE, MSE, RMSE, RMSLE.

The R Squared of each model explains on how accurately the other variables are explaining about the target variable (Ultimate Incurred Insurance Claim Prediction).

MAE is calculated by averaging the absolute differences between actual and predicted Insurance Claim values.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Figure 2

N = the count of the data points.
Y = the actual value in the dataset
Y cap = the model's predicted value.

Fig : MAE Formula

MSE is the mean of the squared difference between the dataset's actual value and the model's predicted Insurance Claim values.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Figure 3

N = the count of data points in the data.

Y = the actual value in the dataset.

Y cap= the model's predicted value.

Fig : MSE Formula

The R.M.S error is calculated using the square root of the average of the squared difference between the prediction and the actual number. The sample standard deviation of the discrepancies between anticipated and observed values is shown below. The following formula is used to compute it:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

Fig : RMSE Formula

Gowthami Nagappan
Teesside University - B1326237@live.tees.ac.uk

The ‘RMSLE’ is predicted log-transformation of RMSE (Root Mean Squared Error) with log transformed actual values.

MODEL PERFORMANCE EVALUATION

Comparison of Each Model Performance

After building a Linear Regression / Decision Tree / Random Forest / SVM model, the Python regression metrics was used to calculate the accuracy of the model prediction and the results are shown below

Evaluation Metrics	Linear	Decision Tree	Random Forest	Support Vector Machine
R^2	0.68	0.51	0.72	0.19
MAE	1330	1485	1151	2148
MSE	3625104	5551583	3155375	9248580
RMSE	1903	2356	1776	3041
RMSLE	7.5	7.7	7.4	8.01

Fig : Evaluation Metrics of all Regression Model

This section examines and compares the results of each regression to discover the optimal model for making highly accurate predictions. R Squared, MAE, MSE, RMSE and RMSLE were used to evaluate each model in this study. Then we looked at all of the regression models to understand how RF was chosen as the efficient model for predicting Insurance Claims.

The evaluation of all ML regression utilized is shown in the graph above. For all ML models, the range of accuracy values was less than 72 percent. The best model was RF, which had a high accuracy of 72.6 percent. According to the findings, RF will effectively address claim prediction issues. Apart from RF, only linear regression provides a higher level of accuracy.

Gowthami Nagappan
Teesside University - B1326237@live.tees.ac.uk

We can see from the above table that RF has the maximum sensitivity which shows that it has a lower error rate than other models which has higher error rates. The model with the lowest RMSE is much better than others. The results indicate an error rate of 1776 and a log transformed error rate of 7.48 which shows that RF has a good level of accuracy.

RESULT ANALYSIS

Performance of Random Forest based on New Data Insurance Claim Prediction

The accuracy of R.F Regression model was assessed using data from various features. The training and validation data for the models show that there is a lower Error rate with high accuracy.

The Random Forest Regression was used to predict ultimate insurance claim by using Test dataset which had 36,000 new records. We predicted a total of 36,000 Insurance claim values using Random Forest Model and Saved in new csv file called submit.csv as shown in image below.

Submit.csv ×

1 to 25 of 36000 entries Filter

ClaimNumber	UltimateIncurredClaimCost
WC8145235	846.53527419
WC2005111	1022.43051588
WC6899143	872.05163966
WC5502023	268.87477408000007
WC4785156	451.77945402000006
WC7296554	1411.4025768200001
WC9274579	1993.94149909
WC7580073	950.69608262
WC5186430	453.20942852999997
WC8599800	975.1714616300002
WC8714609	2579.0890976
WC7379178	1097.79225363
WC8450357	1971.2494300600003
WC3308035	637.2739521599999
WC4536177	751.7868252
WC8699463	4556.264588159999
WC5549096	622.6188560300001
WC1635536	615.2390627300001
WC6913326	1764.5779456400003
WC9531491	856.0127587999999
WC7455291	877.1948421700001
WC4057459	457.8066331900001
WC7798575	1004.6845784700001
WC7590457	3117.09543262
WC1632165	560.56926709

Show 25 per page 1 2 10 100 1000 1400 1430 1440

Fig : Result from Test Dataset Prediction

Gowthami Nagappan
Teesside University - B1326237@live.tees.ac.uk

ETHICAL CONCERNS

- An A.I model can be biased when the decision can be prejudiced against certain cases.
- Lack of transparency as the underlying logic of systems is even hidden from the developers which pose the most ethical and legal concerns.
- Privacy and data protection is also a major factor causing ethical issues. The insurance claiming systems need to be compliant with the latest regulations of the respective country.
- In Auto Prediction, security is also a serious ethical concern. Because of the system's complexities, hackers may try to introduce inaccurate data into it, affecting the system's decision-making output.

DISCUSSION AND CONCLUSION

The findings suggest that the Random Forest model could be used to predict the intensity of claims which is a regression problem in machine learning. The R.F model can achieve the accuracy level of 0.72 only by employing all the features. As a result, Random Forest is a scalable strategy, especially in terms of the number of characteristics. As a result, the model can be utilized for handling large-scale big data challenges. Comparing with the other 3 models, RF has low error rate and 72% accuracy which are good to predict the insurance claims. Given the current and limited resources, the regression RF model has proven to be the most efficient model for prediction among the other methods which we have examined.

FUTURE WORK

Comparative studies while using dataset for different ML and deep learning models would be beneficial. It would also be useful to repeat this analysis using a different insurance branch dataset to see if random forests are still the most accurate predictor. The model can be still evaluated by using various techniques like parameter tuning and parallel computing of model to improve accuracy and performance.

Gowthami Nagappan
Teesside University - B1326237@live.tees.ac.uk

REFERENCE

[1] Weerasinghe, K. P. M. L. P., and M. C. Wijegunasekara. 2016. A comparative study of data mining algorithms in the prediction of auto insurance claims. European International Journal of Science and Technology 5: 47–54.

[2] Jing, Longhao, Wenjing Zhao, Karthik Sharma, and Runhua Feng. 2018. Research on Probability-based Learning Application on Car Insurance Data. In 2017 4th International Conference on Machinery, Materials and Computer (MACMC 2017). Amsterdam: Atlantis Press.

[3] Kowshalya, G., and M. Nandhini. 2018. Predicting fraudulent claims in automobile insurance. In Proceedings of the 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, April 20–21; pp. 1338–43.

[4] Dewi, Kartika Chandra, Hendri Murfi, and Sarini Abdullah. 2019. Analysis Accuracy of Random Forest Model for Big Data—A Case Study of Claim Severity Prediction in Car Insurance. Paper presented at 2019 5th International Conference on Science in Information Technology (ICSITech), Yogyakarta, Indonesia, October 23–24; pp. 60–65.

[5] Singh, Ranjodh, Meghna P. Ayyar, Tata Venkata Sri Pavan, Sandeep Gosain, and Rajiv Ratn Shah. 2019. Automating Car Insurance Claims Using Deep Learning Techniques. Paper presented at 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), Singapore, September 11–13; pp. 199–207.

[6] Stucki, Oskar. 2019. Predicting the Customer Churn with Machine Learning Methods: Case: Private Insurance Customer Data. Master's dissertation, LUT University, Lappeenranta, Finland.

[7] Abdelhadi, Shady, Khaled Elbahnasy, and Mohamed Abdelsalam. 2020. A proposed model to predict auto insurance claims using machine learning techniques. Journal of Theoretical and Applied Information Technology 98: 3428–3437.

[8] Pesantez-Narvaez, Jessica, Montserrat Guillen, and Manuela Alcañiz. 2019. Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. Risks 7: 70.

[9] Sabbeh, Sahar F. 2018. Machine-learning techniques for customer retention: A comparative study. International Journal of Advanced Computer Science and Applications 9: 273–81.

[10] Shalev-Shwartz, S.; Ben-David, S. Understanding Machine Learning: From Theory to Algorithms; Cambridge University Press: Cambridge, UK, 2014. 20.

[11] Huang, S.; Cai, N.; Pacheco, P.P.; Narrandes, S.; Wang, Y.; Xu, W. Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genom.-Proteom. 2018, 15, 41–51.

Gowthami Nagappan
Teesside University - B1326237@live.tees.ac.uk

APPENDIX

Dataset Summary

ClaimNumber : Unique policy identifier
DateTimeOfAccident : Date and time of accident
DateReported : Date when the accident reported
Age : Policy Holder Age
Gender : Gender of the Policyholder
MaritalStatus : Martial status of policyholder. M - Married, S - Single, U - Unknown.
DependentChildren : The no of dependent children
DependentsOther : The no of other dependents without children
WeeklyWages : Total weekly wages
PartTimeFullTime : Binary P or F
HoursWorkedPerWeek : Total number of hours worked per week
DaysWorkedPerWeek : Total number of days worked per week
ClaimDescription : Text description of the accidents and claim
InitialIncurredClaimCost : Initial estimate cost by the insurer on claim cost
UltimateIncurredClaimCost : Total claims payments made by the insurance firm.

Gowthami Nagappan
Teesside University - B1326237@live.tees.ac.uk