Capstone Project: Regression model of Auto Insurance claims

Jean Paul Ngezigihe

Valentine Udeh

Toronto, ON, CANADA

Vector Institute



November 24th, 2022

Abstract

Claim prediction is very essential in the insurance sector as it helps insurance companies to recommend the right type of policy to a prospective policyholder to minimize fraud. This is very necessary especially now that the number of car owners are on the rise and prospective customers present incomplete or missing information. Thus, this study was conducted to determine a powerful machine learning technique that can predict auto insurance claims of a customer with available customer information. A large dataset with customer information and corresponding claim amount on Kaggle was subjected to linear regression and decision tree regression machine learning analysis on google colab. The model evaluation metrics of the two models showed that the decision tree regression model had a higher coefficient of determination (77%) than the linear regression model (66%).

# 

# 

# 

# 

# 

[**INTRODUCTION**](#_n0mtwg4rsns) **2**

[**Problem Definition**](#_75rf4vta81ax) **3**

[Brief on dataset](#_5fzus8rfoqyt) 3

[Problem you are trying to solve](#_oeus9e19087) 4

[Proposed model and approach](#_tzpceeahihq0) 4

[**Data exploration and description**](#_lquiyrwpy6ke) **4**

[Data splitting](#_9lc5qclmccbg) 7

[**Model ( Describe your model and algorithm)**](#_kxv9w3r2kjp) **8**

[**Results and findings**](#_h44yr4n0dohp) **10**

[Interpretation](#_bam54e4nwcpn) 10

[**CONCLUSION**](#_chou9188p6co) **11**

[**REFERENCES**](#_ksfxh7wzck1c) **12**

# 

# INTRODUCTION

Auto insurance simply refers to a contract between a car owner and an insurance company which protects the customer from financial loss in the case of theft or accident. In most developed countries like Canada, all motorists are legally required to be auto insured. Thus, with the geometric increase in the number of car owners, there has been a tremendous increase in the number of auto insurance claims being registered (Singh et al., 2019). For instance, in the United States, accident claim severity increased by 35% between 2010 and 2019 leading to a significant increase in insurance expenditure (Hanafy and Ming, 2021). The life cycle of registering, processing and making a decision for each claim involves the manual examination by the service engineer who creates the damage report followed by the physical inspection by a surveyor from the insurance company which makes it a long-drawn-out process (Singh et al., 2019).

Claim prediction is an important operation in the insurance sector as this helps insurance companies to recommend the right type of policy to a prospective policyholder (Abdelhadi et al., 2020). Insurance rates are dependent on lots of factors like age, gender, location, employment status, marital status, income, vehicle type and many others. There are meta datasets on auto insurance claims in reputable repositories like Kaggle that can be used to develop predictive models using machine learning. These models could be used to predict insurance claims even when there is lots of missing information from a proposed policy holder. Many studies have been conducted using various machine learning approach: linear regression, logistic regression, decision tree, XGBoost, random forest, naïve Bayes, and neural networks to develop auto insurance predictive models (Hanafy and Ming 2021; Abdelhadi et al., 2020; Singh et al., 2019). However, the development of an optimal predictive model for total claim amount is yet to be fully addressed.

Thus, this study is geared towards investigating more powerful machine learning techniques to predict auto insurance claims by using a large dataset from Kaggle.

# Problem Definition

## Brief on dataset

The dataset used in this study is the “[Auto insurance](https://www.kaggle.com/datasets/ranja7/vehicle-insurance-customer-data)” data from Kaggle. It consist of 23 attributes, which includes 22 input features: State,Response, Coverage, Education, Effective To Date,'EmploymentStatus, Gender, Location\_Code, Marital\_Status, Policy\_Type, Policy, Renew Offer Type,Sales Channel, Vehicle\_Size, Vehicle\_Class, Customer Lifetime Value, Income, Months Since Last Claim,Monthly Premium Auto, Months Since Policy Inception, Number of Open Complaints and Number of Policies; and 1 output variable: Total Claim Amount.

## Problem you are trying to solve

The capstone problem was to train model(s) that can predict the total claim amount using the input features.

## Proposed model and approach

The study was carried out using standard regression models (Linear Regression and Decision Tree) to develop the prediction model.

# Data exploration and description

According to Kaggle, this dataset contains customer data with their vehicle insurance policies. Details about customers and the insurance taken for their vehicles are provided which can be explored to segment similar kinds of customers.

The data contains both categorical and numerical variables. It contains useful socio-economic information about the customer such as income, marital status, location, etc. These attributes are important to understand the customer behavior. These attributes will be used as features in my model, in other words, “X variable”. The target, in other words, the “Y variable”, will be the total claim amount column, which is what we are trying to predict.

We started by exploring the dataset and dimensions. We did so by doing a summary statistic for each variable. Then, we determined if there are any missing data points in the dataset. This is a typical way to gain an initial understanding of a dataset.

Our dataset consists of 9134 rows and 23 columns. It is a mix of numerical, float and integers. There were no missing data points so we did not need to use data imputation strategies to substitute the missing values. However, there were some 163 duplicated rows so we had to eliminate them.

In addition to the descriptive statistics, we visualized the data to provide additional information on the data itself and can guide us in how to carry out the analysis appropriately.

**Table 1: Descriptive statistics of some of the numeric variables in the dataset**

|  | **Customer Lifetime Value** | **Income** | **Monthly Premium Auto** | **Months Since Last Claim** | | **Months Since Policy Inception** | | **Number of Open Complaints** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **count** | 9134 | 9134.00 | 9134.00 | 9134.00 | | 9134.00 | | 9134.00 |
| **mean** | 8004.940 | 37657.38 | 93.22 | 15.10 | | 48.06 | | 0.38 |
| **std** | 6870.968 | 30379.90 | 34.41 | 10.07 | | 27.91 | | 0.91 |
| **min** | 1898.008 | 0.00 | 61.00 | 0.00 | | 0.00 | | 0.00 |
| **25%** | 3994.252 | 0.00 | 68.00 | 6.00 | | 24.00 | | 0.00 |
| **50%** | 5780.182 | 33889.50 | 83.00 | 14.00 | | 48.00 | | 0.00 |
| **75%** | 8962.167 | 62320.00 | 109.00 | 23.00 | | 71.00 | | 0.00 |
| **max** | 83325.381 | 99981.00 | 298.00 | 35.00 | | 99.00 | | 5.00 |

The data type and count of all the features and the target variables is presented below:

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 State 8971 non-null object

1 Customer Lifetime Value 8971 non-null float64

2 Response 8971 non-null object

3 Coverage 8971 non-null object

4 Education 8971 non-null object

5 Effective To Date 8971 non-null object

6 EmploymentStatus 8971 non-null object

7 Gender 8971 non-null object

8 Income 8971 non-null int64

9 Location Code 8971 non-null object

10 Marital Status 8971 non-null object

11 Monthly Premium Auto 8971 non-null int64

12 Months Since Last Claim 8971 non-null int64

13 Months Since Policy Inception 8971 non-null int64

14 Number of Open Complaints 8971 non-null int64

15 Number of Policies 8971 non-null int64

16 Policy Type 8971 non-null object

17 Policy 8971 non-null object

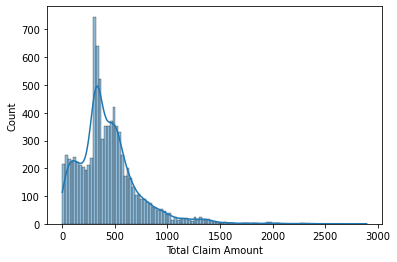
18 Renew Offer Type 8971 non-null object

19 Sales Channel 8971 non-null object

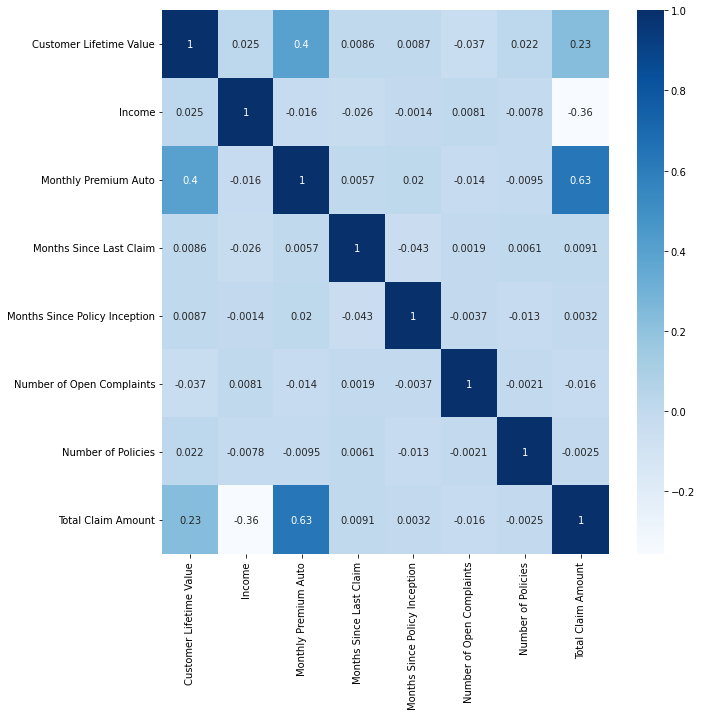
20 Total Claim Amount 8971 non-null float64

21 Vehicle Class 8971 non-null object

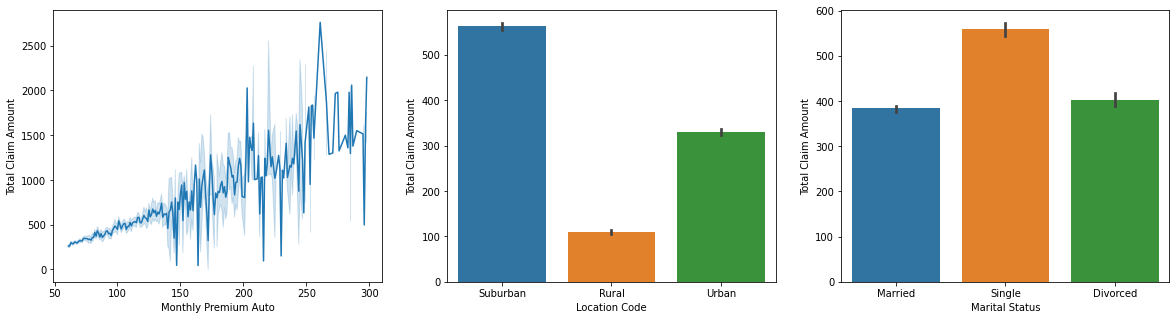
22 Vehicle Size 8971 non-null object



Our data appears to be rightly skewed. We transformed the data into a gaussian distribution. Then we used a heat map to see which variable has high correlation.



We can see that “Monthly Premium Auto” & “Customer lifetime value” has the highest correlation with our target variable “Total Claim Amount”. Other variables seem to have low correlation . We then plotted the relationship between our features and targets.



As foreseen, insurance claim amounts tend to increase as a function of Monthly premiums auto. We also see the same trend with customers in the suburban location code. On the other hand, it is difficult to draw any conclusion regarding marital status.

**Feature encoding**

Some of the features are non-numeric, we must transform them via feature encoding, since we wish to use a regression model. We will be using a simple label encoder that turns the target labels to values between 0 and n\_labelclasses - 1. We will then remove the columns with non-numeric values with the new encoded ones.

Now that we have explored our data and did some preprocessing, we are now ready to create our model.

## Data splitting

As mentioned, we will build a model to predict the amount of insurance claims based on customer attributes such as income, marital status and what region they live in. We have split the dataset into different datasets: (1) training set; and (2) test set. The dataset is split into 70% training set and 30% test set. See Part 3 of Code for reference.

# 

# 

# 

# 

# 

# 

# 

# 

# Model

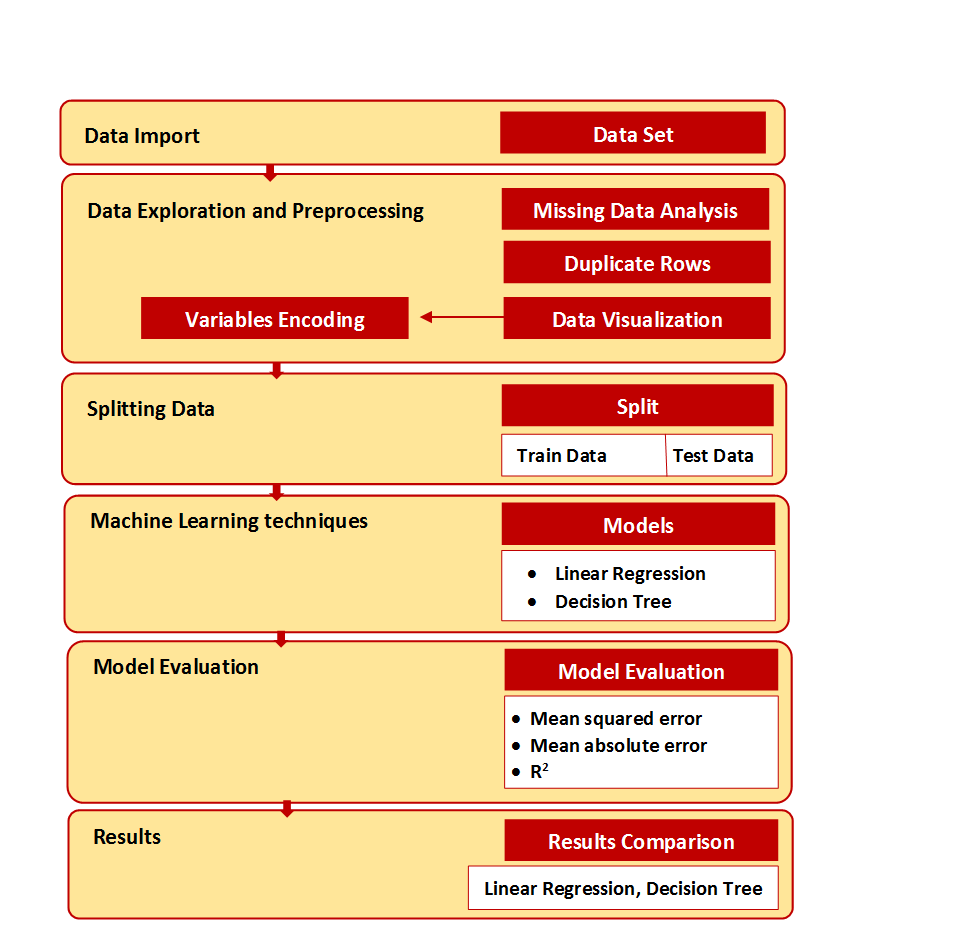


Fig 4. Overall Structure of the study model

**Models:**

1. **Linear regression model:** Linear regression is one of the well-known algorithm in machine learning that is used to estimate the linear relationship between one or more features (predictor variables) and a continuous target (response) variable. For this study our target variable (Total claim amount) is a continuous variable so this model is appropriate for it.

A basic linear regression model: **y = wTx + b**

**where :** x = (x1, …., xD) is the features; y is the target; w is the weights of the parameters and b is the bias/intercept.

1. **Decision tree regression model:** A decision tree is a supervised learning approach used to solve classification and regression issues, but it is mostly used to solve classification issues. It is a classifier organized by the tree structure, where the internal nodes are the data variables, the branches are the decision rules, and each node is the output. It consists of two nodes. One of them is a decision node used for decision-making, and it has various branches. The second node is a leaf node, which represents the result of these decisions. In this study, we used decision tree regression because our target (total claim amount) is a continuous variable.

**Model Evaluation**

There are common metrics that are used to evaluate the performance of a linear regression model include Mean Squared Error (MSE), Symmetric mean absolute percentage error (SMAPE), Mean Absolute Error (MAE) and the coefficient of determination (R^2). The R squared metric has many advantages over MSE and MAE. For instance, it is more informative (Chicco, D., Warrens, M. J., & Jurman, G. , 2021). The positive values of the coefficient of determination range in the [0, 1] interval, with 1 meaning perfect prediction. On the other hand, the values of SMAPE range in the [0, 2], with 0 meaning perfect prediction and 2 meaning worst prediction possible (Chicco, D., Warrens, M. J., & Jurman, G. , 2021). This can be confusing for interpretability. Therefore, R^2 is chosen. In simple terms, the coefficient of determination ( R^2 ) is a measure of how well the model fits the dependent variable.

# 

# 

# 

# 

# Results and findings

In the introduction, we discussed our goal to introduce more powerful machine learning techniques to predict auto insurance claims. After building our model, this section will discuss the success of our hypothesis. Table 2 below presents the summary of our model metrics.

**Table 2: Evaluation metrics of the models**

| Metrics | Linear Regression | Decision Tree Regression |
| --- | --- | --- |
| **Mean squared error (MSE)** | 54.604 | 32.531 |
| **Mean absolute error (MAE** | 5.805 | 4.393 |
| **Coefficient of determination R2** | 0.609 (61%) | 0.767 (77%) |

The linear regression model had R2 of 61% while the decision tree model had an R2 value of 77%. The relatively high R2 value of the Decision Tree model indicates that the model can explain 77% of the variance in the data, so it performs quite well!

## Interpretation

The results of this model show that by taking specific attributes of an individual such as Policy, Type, Renew Offer Type,Sales Channel, Vehicle Size, Vehicle Class, Months Since Last Claim,Monthly Premium Auto, we can predict the approximate amount of medical costs (based on insurance claims) the individual is likely to claim.

Although not perfect, this model can be useful for insurance companies in many ways as discussed in the introduction. We believe companies can use it in tandem with existing processes to manage future cash flows as they can predict an estimate of future outgoing cash flows. Additionally, they can create new insurance policy products targeting specific customer groups. Furthermore, companies can reduce business risk by reducing the onboarding of “risky” customers with certain attributes. However, there are growing concerns of companies using machine learning to further systematic oppression on vulnerable communities (Aniek F. Markus, Jan A. Kors, Peter R. Rijnbeek, , 2021).

To go further, we could implement partial dependence plots to increase explainability of our model by showing the dependence of the predicted response on a single feature. To assess fairness, we can do subpopulation analysis which assess whether our model behaves identically across these populations. In combination, these techniques would make our model more ethically responsible and reduce possible harm from our model.

# CONCLUSION

In conclusion, throughout this study, by using the Kaggle data set “Auto Insurance”, we were able to predict insurance claim amounts. Through data exploration, we investigated our dataset which gave us insights on the data. Then, we performed some data cleaning and preprocessing. This made our dataset ready to build our model. The regression model had a R square score of 61%, while the decision tree had 77%. We can infer that the non-linear model was a better model for this problem. We are confident our study showed machine learning can be useful for insurance companies, and our model is a good example.

However, our model can be improved. The accuracy can be improved by tuning certain parameters. Additionally, there were many varales with low correlation. We could recreate the model by excluding some low importance features. In future, we can also implement an ethical framework to ensure it is used accordingly by insurance companies.

# 

# 

# 

# 

# REFERENCES

Abdelhadi, S., Elbahnasy, K., & Abdelsalam, M. (2020). A proposed model to predict auto insurance claims using machine learning techniques. *Journal of Theoretical and Applied Information Technology*, *98*(22), 3428–3437.

Aniek F. Markus, Jan A. Kors, Peter R. Rijnbeek, (2021).The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies,Journal of Biomedical Informatics,Volume 113,2021,103655,ISSN 1532-0464,https://doi.org/10.1016/j.jbi.2020.103655.(https://www.sciencedirect.com/science/article/pii/S1532046420302835).

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ. Computer science, 7, e623. https://doi.org/10.7717/peerj-cs.623

Fauzan, M. A., and Murfi, H. (2018). The accuracy of XGBoost for insurance claim prediction. Int. J. Adv. Soft Comput. Appl, 10(2).

Paruchuri, H. (2020). The Impact of Machine Learning on the Future of Insurance Industry. American Journal of Trade and Policy, 7(3), 85-90.

SARKAR, R. (2019, May 31). Insurance. Kaggle. <https://www.kaggle.com/datasets/ranja7/vehicle-insurance-customer-data>

Singh R., Ayyar M.P., Sri Pavan T.V., Gosain S. and Shah R.R., "Automating Car Insurance Claims Using Deep Learning Techniques," 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), 2019, pp. 199-207, doi: 10.1109/BigMM.2019.00-25.

Stanley, K. D., Gris, M., & Anderson, J. M. (2020). *Autonomous Vehicles and the Future of Auto Insurance*.<https://doi.org/10.7249/RRA878-1>

Wang, H. D. (2020). Research on the Features of Car Insurance Data Based on Machine Learning. *Procedia Computer Science*, *166*, 582–587. [https://doi.org/https://doi.org/10.1016/j.procs.2020.02.016](https://doi.org/https:/doi.org/10.1016/j.procs.2020.02.016)