Data Pre-processing: Case Study on Car Insurance Dataset

Jia Hui Tang[1], Ka Li Lim[1] and Yu Jie Chong[1]

[1*]Department, Faculty of Information Science and Technology, Bangi, 43600, Selangor, Malaysia.

Contributing authors: a176297@siswa.ukm.edu.my; a176496@siswa.ukm.edu.my; a176730@siswa.ukm.edu.my;

Abstract

The ability to effectively predict and identify the circumstances in which a submitted claim of insurance by policyholders should be approved by the insurer or otherwise is crucial for the benefit of both parties. Insurers that establish accurate claims reserves not only reduce their own financial losses due to inaccurate value of claims and being unaware of low insurance reserves, whilst fulfilling their financial obligations without unnecessary delays in insurance settlements of the insured individuals. Nevertheless, such ability is a stumbling block faced by insurance companies as they require highly experienced staff with a keen understanding of policyholders' personal and driving-related backgrounds before manually approving insurance claims as submitted. This quickly becomes infeasible when the insurance claim submissions or data surge, which indicates the increase in demand for experienced manpower to conduct the approval of insurance without delays. This justifies the need for insurers to start adopting and implementing machine learning algorithms to enhance the interpretation of insurance approvals without the excessive demand for experienced manpower and time. Before the development and deployment of a machine learning predictive model, data pre-processing is one of the crucial stages in obtaining high-quality data, which is crucial for machine learning models to learn and maximize the utilization of hidden information in the data. Therefore, this project intends to explore different data pre-processing techniques on a car insurance dataset and the discussions are segmented into Exploratory Data Analysis (EDA), data cleaning and data reduction. Firstly, an overview of the characteristics of dependent and independent attributes is explored and presented in the form of data quality reports, histograms, correlation heatmap and boxplots. Next, several data cleaning techniques are explored such as handling missing values via simple and iterative imputation, data scaling with Standard Scaler, data discretization with K bin Discretizer, feature selection with SelectKBest, SelectFromModel, Recursive Feature Selection with Cross-Validation, and Sequential Feature Selector. Data pre-processing ends by exploring dimensionality reduction

with PCA and Feature Agglomeration. Upon data pre-processing, data distribution of the cleaned data, which is of better quality is visualized via histograms and presented in order to picture the effect of data pre-processing on raw data.

Keywords: Data Pre-processing, Data Exploration, Data Cleaning, Data Reduction, Data Visualization, Car Insurance Claim

1 Introduction

As we work through our daily task day by day, risk and uncertainty are inevitable to our lives. The possibility and thoughts of suffering from an accident or natural disasters has made us to feel insecure. With that being said, insurance is known to plays an important role in helping us to cope through that feeling. Generally, an insurance is a form of protection that help can manage our risk by providing financial compensation. Thus, in an event where an accident has occurred and a hefty payment is needed, the policyholder can file an insurance claim to their policy provider to compensate the amount of loss involved.

As a matter of fact, with the progressive growth of novice drivers throughout the years, it is almost assured that the occurrence of insurance claim for the coverage of property damage or medical treatment related to automobile accident might as well increase. Indirectly, this also means the increment of insurance being purchased by vehicle owner in order to reduce their loss that may occur in lights of an accident taking place. However, as the rise in insurance application might be able to make the insurance business more thriving, this also may attract certain individual to commit insurance fraud. For that reason, the insurance company or policy provider must ensure that certain requirement must be made being approving these insurance claim. By following these set of rules, the company will then be able to prevent any financial issue whilst upholding their end of the contract.

Henceforth, it important to take into account of all the factor that might influence a policyholder to file a claim against the insurance, and this can be accomplished by having a deep understanding towards the needs of our customer. In the insurance field, it is crucial to assess the level of risk that each policyholder may poses to the company [1] because each client is unique, and also the insurers need to anticipate the losses precisely so that they can set rates that reflect the risks they are covering. During the assessment, the risk analyst expert will look through specific attribute and characteristic that might pose a problem toward the premium insurance policy. Furthermore, as mentioned in the Auto Insurance Guide, there are myriad of aspect that may influence the car insurance premiums.[2] In particular, there are a few elements that stand out among the rest such as the customer's credit rating, driving record and age. For instance, the credit rating is very important in determining the credibility of a

policyholder because it will be able to provide us with insights on how competently the customer perform on their financial management. So, with information such as credit card payment, outstanding debt, and others, we can then make an informative decision on what their credit rating will be as this may influence the premium insurance that will be offered. It is important to perform this assessment on policyholder because determining the level of risk each policyholder carries will be able to help us classify into the categories of insurance premium they belong to, so the higher the risk they have, the more expensive their insurance coverage will be and vice versa. Therefore, in this way, we will be able to provide an appropriate or fair insurance cost by customizing each insurance premium based on their traits without overcharging any of the clients.

Nevertheless, with the advancement of technology at the moment, the application of data mining on insurance claim data is being widely implemented. Various machine learning approach have been proven to be useful in processing a large amount of data in order to extract valuable information which will be able to help during decision making process. Thus, by developing a classifier model to predict insurance claim, we will then be able to ensure a smoother insurance registration process for all the vehicle owners. However, in order to build the best machine learning model with high accuracy and efficiency, there are a few initiatives that we must go through first and one of them is data pre-processing. Data pre-processing is considered a critical step in creating a machine learning model because it can help to improves the data quality and also facilitates the extraction of relevant insights from the data.[3] The process involves the technique to prepare the raw data by cleaning and organizing it in order to transform the data into a format which will be understandable and readable by the machine learning model. As a result, because the quality of data can impair the model's ability to learn, it is important that we pre-process the data before feeding it into the model. In addition to that, there are also certain studies that have been done that show the implementation of pre-processing method have drastically improved the performance of the classification algorithm.[4] In short, the few data pre-processing that will be done in this study will include data exploration, data cleaning and data reduction. Other than that, the initial stage in machine learning data preparation is to acquire the dataset. So, in this case, the dataset that we will be using is the Car Insurance Data which contain data about the customer behaviors which are relevant for its annual car insurance. The outcome column in this dataset has also indicates on whether the customer have claimed their insurance. Consequently, this dataset will be suitable in determining whether a customer is likely to file an insurance claim over their car.

In conclusion, the main objective of this project is to perform data pre-processing technique onto the chosen dataset which is the Car Insurance Data in order to prepare a better-quality data that can be used to train the model later

on. The data pre-processing step that will be taken in this procedure will include listing of attribute with their types, data visualization, handling missing values, noise, inconsistent data, data transformation, feature selection and dimensional reduction. For the handling of missing value, there will be two approach that will be applied on the data which are the univariate and multivariate approach. Meanwhile in term of feature selection, four different methods known as the SelectKBest, SelectFromModel, Recursive Feature Selection with Cross Validation and Sequential Feature Selector will first be performed on the dataset and then the method that return the most appropriate relevant feature which will be able to improve the performance of the model will then be chosen. Finally, two different dimensionality reduction method, Principal component analysis (PCA) and Feature Agglomeration is then used and the one that provided a better result among both of them will then be picked. Data pre-processing plays an important role in delivering a quality data for the machine learning model; therefore, it is undeniable that we must not treat this process lightly and with utmost importance.

2 Related Work

As the implementation of machine learning continue to evolve in parallel with the rise of artificial intelligence, insurance company are also rapidly embracing the application of machine learning model in order to achieve critical outcomes such as cost reduction, improved underwriting, and make accurate insurance claim prediction. In the insurer's point of view, this a wise business strategy because by exploring useful information that have been collected, we can then gain more insight about our policyholder and in return use that knowledge to build a client profile which will be able to help us in determine whether an insurance claim will be filed during the period they are insured. Nonetheless, there have been several studies that was done throughout the years that have discuss about the prediction of insurance claim with the application of machine learning. In particular, research that focus on the machine learning approaches for auto insurance big data has been done.[5] In this paper, it is shown that the researcher, Hanafy and Ruixing has utilized several machine learning models such as Logistic Regression, XGBoost, Random Forest, Decision Trees, Naive Bayes and K-NN in order to assess the prediction on the occurrence of insurance claim. Other than that, the data pre-processing steps that was taken in this research also includes visualizing the distribution of the target variable, handling missing value, correlation overview, and implementation of feature. So, based on the evaluation on the performance of the model, it is shown that Random Forest is the better model among the other.

Apart from that, there are also other studies such as the research done by Shady, Khaled, and Abdelsalam in 2020 whereby they have proposed a model that is able to predict auto insurance claim by using machine learning

technique.[6] In their studies, they have focus on using machine learning algorithm such as Artificial Neural Network (ANN), Decision Tree (DT), Na¨ıve Bayes and XGBoost to build a precise model to predict car insurance claims. Furthermore, the data pre-processing that was used to improve the predictive effect are as such, discarding irrelevant data, handling missing data, variables discretization, variables encoding, and standardization. As a result, the result that their studies gotten showed that the XGBoost and Decision Tree has achieved the best accuracy among the other models which is 92.53 percent and 92.22 percent respectively. Besides that, a study done by Endalew and Teklu in 2021 about motor insurance claim status prediction using machine learning techniques such as Random Forest (RF) and Support Vector Machine (SVM has shown that the RF classifier is slightly better than SVM.[7] Their findings is also proven by the fact that the RF classifier has a higher accuracy at 98.36 percent as compared to SVM at 98.17 percent. Also, the data pre-processing technique that was used for the data set preparation also include, data cleaning where noisy and irrelevant data is removed, data integration, data normalization or data transformation where z-score is being used and lastly data encoding specifically one-hot encoding to convert categorical data into numeric or binary.

Additionally, Huangfu has also published a paper related to data mining for car insurance claim prediction in 2015 whereby several machine learning algorithms has been used.[8] In this study, the machine learning algorithm that was being implemented was Logistic Regression, The Tweedie Model, Regression Tree, The Conditional Inference Decision Tree, GLM Gamma Regression, SVM and K-means Clustering and based on the result obtained, it seems that the combination between PCA with a Regression Tree has produced the best result. Although the data pre-processing step was not specified in this paper, but the researcher has pointed out that the main method used to deal with the high dimensionality of the data set is through Principal Component Analysis (PCA) and Response Averaging, and then later on evaluate the model using the normalized Gini coefficient. Also, in 2021, a study has been done on the application of machine learning and data visualization in the insurance sector with the main purpose to identify meaningful and decisive pattern for claim filing.[9] On the contrary, while other studies might only focus their analysis with only one data set, the researcher on these studies has decided to perform claim analysis using ML classification algorithm on two different data set. However, apart from that, the steps taken in other procedure is quite similar to other studies. For example, the data preparation method that is involved in this research is none other than data cleaning, exploratory data analysis (EDA), normalization and dimensionality reduction processes. In short, for both the data set, Random Forest is proven to be the classifier with suitable feature selection methods.

Moreover, a thesis title as “A Framework to Forecast Insurance Claims” written by Tim Pijl has also discussed about the possibility of improving forecast

upcoming claim prediction based on previous client data.[10] In this thesis, Tim have decided to implement four different classification technique which are Decision Tree, Random Forest, Binary Logistic Regression, and a Support Vector Machine algorithm. The classification algorithm will then assign a probability will be able to help in predicting whether a claim will be filed. In this research, the data preparation process that are involved include feature generation, exploratory analysis, data cleaning which consist of removing missing value or outlier, and dimensionality reduction. For dimensionality reduction, there are three different techniques that are used: variable elimination, application of random forest to obtain informative variable and multiple correspondence analysis. So, depending on whether to model a data-driven decision or variable influence driven decision, the model that was recommended varies, in which that random forest or SVM is a more suitable for the former while a decision tree is better for the later as it can be used without too much loss if accuracy. Other studies that are done also include the research done the features of car insurance data based on machine learning. [11] In this paper, the Random Forest, Gradient Lifting Tree (GBDT) and Lifting Machine Algorithm (LightGBM) is being compared and the one that managed to stand out among the three is LightGBM that has the best superiority and robustness. Other than that, the data pre-processing involve are simple data cleaning such as handling missing value and remove redundant feature, feature renaming, and eigenvalue processing.

In general, it seems that although the data pro-processing steps taken for each study is unique and different, however it is undeniable that this process is important as it provides a solid foundation for model development. Based on all the studies above, we can see that most of the studies have implemented a suitable pre-processing technique that is appropriate according to the data set used. Therefore, the main purpose of this project is to divulge about the importance of data pre-processing and then select the best technique that can bring out the most insightful information which in turn will later be used to feed into machine learning model that will help us with the insurance claim prediction. Several techniques that are already mentioned above and newer one will then be implemented and compared in order to search for the best data pre-processing method.

Table 1. Overview on different studies that is done on similar dataset

Author	Article	ML Algorithms	Data Pre-processing Techniques
Dan Huangfu	Data Mining for Car Insurance Claims Prediction (2015)	Logistic Regression, The Tweedie Model, Regression Tree, The Conditional Inference Decision Tree, GLM Gamma Regression, SVM and K-means Clustering	Exploratory Data Analysis (EDA), Principal Component Analysis (PCA), and Response Averaging
Tim Pijl	A Framework to Forecast Insurance Claim (2017)	Decision Tree, Random Forest, Binary Logistic Regression, and Support Vector Machine (SVM)	Feature Generation, Exploratory Data Analysis (EDA), Data Cleaning and Dimensionality Reduction (Variable Elimination, Reduction through Random Forest, Multiple
Hui Dong Wang	Research on the Features of Car Insurance Data Based on Machine Learning (2019)	Random Forest, Grdient Lifting Tree (GBDT), and Lifting Machine Algorithm (LightGBM))	Data Cleaning, Feature Renaming, Eigenvalue Processing
Shady Adelhadi, Khaled Elbahmasy, and Mohamed Adelsalam	A Proposed Model to Predict Auto Insurance Claims using Machine Learning Technique (2020)	Artificial Neural Network (ANN), Decision Tree (DT), Naïve Bayes, and XGBoost	Data Cleaning, Variables Discretization, Variables Encoding, and Data Standardization
Mohamed Hanafy, Ruixing Ming	Machine Learning Approaches for Auto Insurance Data (2021)	Logistic Regression, XGBoost, Random Forest, Decision Tree, Naïve Bayes, and K-NN	Data Visualization, Data Cleaning, Correlation Overview, and Feature Selection
Endalew Alamir, teklu Urgessa, Ashebir Hunegnaw, and Tiruveedula	Motor Insurance Claim Status Prediction using Machine Learning Techniques (2021)	Random Forest, Multi Class-Support Vector Machine (SVM)	Data Cleaning, Data Integration, Data Normalization, and Data Encoding

Seema Rawat, Aakankshu Rawat, Deepak Kumar, and A.Sai Sabitha	Application of machine learning and data visualization techniques for decision support in the insurance sector (2021)	Logistic Regression, Random Forest, Decision Tree, Support Vector Machine (SVM), Gaussian Naïve Bayes, Bernoulli Naïve Bayes, Mixed Naïve Bayes and K-Nearest Neighbours	Data Cleaning, Exploratory Data Analysis (EDA), Data Normalization, Dimensionality Reduction (Chi-Square Test, Recursive Feature Elimination, Tree-based Feature Selection)

3 Material and Methods

The data pre-processing of this research is segmented into 2 main stage – Feasibility study, Data Acquisition and Preparation. The subprocesses of each stage are depicted in Figure 1. A thorough exploration, understanding, analysis, and pre-visualization of the car insurance dataset is performed, followed by several suitable pre-processing techniques to clean and reduce the dataset. The objective of data pre-processing is to ensure data is in an appropriate and feasible format that enable machine learning or deep learning models to learn from data better before producing reliable and accurate prediction on approving car insurance claims made by policyholder.

Table 2. General Workflow of Data Pre-processing Requirements in the Car Insurance Dataset Case Study

Steps	Processes Involved
1: Feasibility Studies	• Identify Business Problem • Determine Business Objective

2: Data Acquisition & Pre-processing	• Data Acquisition and Exploration
	  ○ Data Acquisition: Source
	  ○ Descriptive analysis of data attributes
	  ○ Data quality report
	  ○ Visualize raw data
	• Data Pre-processing
	  ○ Various problems in the raw dataset
	  ○ Data Cleaning: Handling missing values, noise, and inconsistency in dataset

	○ Data Scaling, Data Transformation, Feature Selection and Dimensionality Reduction

3.1 Data Acquisition and Pre-processing

Data Acquisition & Exploration: The car insurance dataset is retrieved from Kaggle at https://www.kaggle.com/datasets/sagnik1511/car-insurance-data whereby it is from a company who shared its annual car insurance data in order to study its real customer or policyholder behavior via the data collected. The aforementioned dataset consists of 18 attributes that resembles practical attributes used by insurance companies in determining the insurance claims behavior or approval to their customers, which is the target variable. There are two possible behaviors portrayed by the customers, which is they had claimed their insurance successfully or otherwise. The dataset is organized into a total of 10, 000 unduplicated insurance claim records with both successful and unsuccessful claims. Each record is made up from 19 columns, 18 of them are real life insurance claims attributes and 1 of them is the target column. Table 2 shows the descriptive features of the dataset which consists of 18 attributes and the target attribute, along with their semantics, characteristics (data type and level) and purpose in the research of car insurance claims. Data quality report for categorical and continuous attributes are presented in Table 3 and Table 4 respectively to record its quality upon thorough analysis of the dataset. The investigation of the dataset is followed by the visualization of the raw data, or data prior pre-processing in order to have a clear picture of the quality of the car insurance data.

3.1.1 Descriptive Analysis

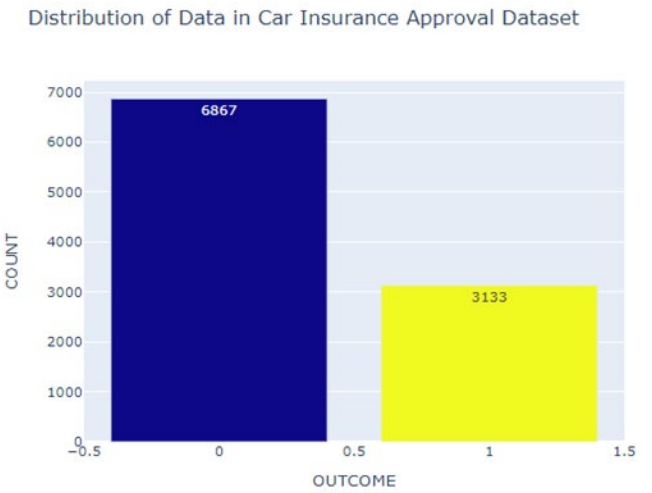


Figure 1 reports on the statistical information of the total number of car insurance claims submitted by car insurance policyholders and received the insurance claim successfully or otherwise. The conclusion drawn is that the majority of the car insurance claims, which are 6867 cases (68.67%) were rejected by the company whereas 3133 of them (31.33%) were accepted. As the number of car insurance claim records disapproved by the insurance company almost twice as higher than the number of approved cases, the dataset is then an imbalanced dataset and the risk of introducing bias in the car insurance claim approval prediction made by the predictive model is high. The justification is the predictive model learns more data pattern or condition in which insurance claim is disapproved than the approved ones.

Table 3. Data attributes and its preliminary information

Attributes/ Features	Data Type	Data Level	Description
Id	int64	Ordinal	Policyholder identification number
Age	object	Interval	Policyholder age (year-old)
Gender	object	Binary	Policyholder gender
Race	object	Categorical	Policyholder race
Driving_Experience	object	Interval	Policyholder driving experience
Education	object	Ordinal	Policyholder education level
Income	object	Ordinal	Policyholder income category
Credit_Score	float64	Numerical	Likelihood of policyholder in filing an insurance claim within the coverage period
Vehicle_Ownership	float64	Binary	Policyholder owns a car
Vehicle_Year	object	Binary	Year of purchase of automobile
Married	float64	Binary	Policyholder marital status (Married)
Children	float64	Binary	Policyholder child / children
Postal_Code	int64	Categorical	Policyholder residential postal code
Annual_Mileage	float64	Numerical	Policyholder automobile annual mileage
Vehicle_Type	object	Binary	Policyholder automobile type
Speeding_Violations	int64	Numerical	The number of speeding violations breached by policyholder
Duis	int64	Categorical	Driving Under Influence (DUI), or referred as DWI (driving while intoxicated), OMVI (operating a motor vehicle impaired) or OVI
Past_Accidents	int64	Numerical	The number of past accidents experienced by policyholder
Outcome	float64	Binary	Permit / Grant of Insurance Claim

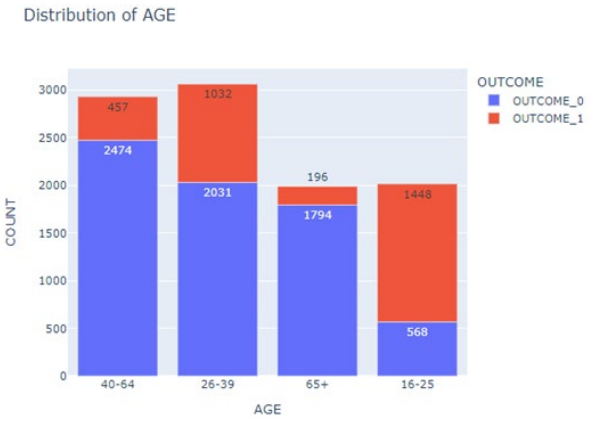
Table 4. Data Quality Report on Continuous Variables

Attributes/ Features	Count	%Missing	Cardinality	Min	25%	Mean	Median	75%	Max	STD
Id	10000	0.00	10000	101.0000	249638.50	500521.91	501777.0	753974.50	999976.00	290030.769
Credit_Score	9018	9.82	9019	0.053358	0.417191	0.515813	0.525033	0.618312	0.960819	0.137688
Vehicle_Ownership	10000	0.00	2	0.000000	0.000000	0.697000	1.000000	1.000000	1.000000	0.459578
Married	10000	0.00	2	0.000000	0.000000	0.498200	0.000000	1.000000	1.000000	0.500022
Children	10000	0.00	2	0.000000	0.000000	0.688800	1.000000	1.000000	1.000000	0.463008
Postal_Code	10000	0.00	4	10238.00	10238.000	19864.548	10238.00	32765.000	92101.00	18915.6139
Annual_Mileage	9043	9.57	22	2000.000	10000.000	11697.003	12000.00	14000.000	22000.00	2818.43453
Speeding_Violations	10000	0.00	21	0.000000	0.000000	1.482900	0.000000	2.000000	22.000000	2.241966
Duis	10000	0.00	7	0.000000	0.000000	0.239200	0.000000	0.000000	6.000000	0.554990
Past_Accidents	10000	0.00	15	0.000000	0.000000	1.056300	0.000000	2.000000	15.000000	1.652454

Table 5. Data Quality Report on Discrete Variables

Attributes/ Features	Count	%Missing	Cardinality	Mode	Mode Frequency	Mode%	2nd Mode	2nd Mode Frequency
Age	10000	0	4	26-39	3063	30.63	40-64	2931
Gender	10000	0	2	Female	5010	50.10	male	4990
Race	10000	0	2	Majority	9012	90.12	minority	988
Driving_Experience	10000	0	4	0-9y	3530	35.30	10-19y	3299
Education	10000	0	3	high school	4157	41.57	university	3928
Income	10000	0	4	upper class	4336	43.36	middle class	2138
Vehicle_Year	10000	0	2	before 2015	6967	69.67	after 2015	3033
Vehicle_Type	10000	0	2	Sedan	9523	95.23	sports car	477

Figure 2. Age Distribution of Car Insurance Policyholders



The age of policyholders ranges from 16 to 65 and above year-old was categorized into 4 groups as depicted in Figure 2. The age group with the most policyholders is the 26–39-year-old group whereas the 65+ year-old group has the least number of car insurance claims submitted to the company. Generally, most of the insurance were not approved by the company, except for age group 16-25 where the proportion of approved claims are higher than disapproved claims. Moreover, the number of car insurance claims submitted is not directly proportional to the number of approved nor disapproved car insurance claims. Even though age group 26-39 has the greatest number of car insurance claims submission compared to age group 16-25, but the number of approved claims (1032 cases) is less than those in age group 16-25 (1448 cases). Nevertheless, an obvious correlation between age and the approval of car insurance claims is the smaller the age of the policyholder, the higher the chance of success to have their car insurance claims approved.

Figure 3. Gender Distribution of Car Insurance Policyholders

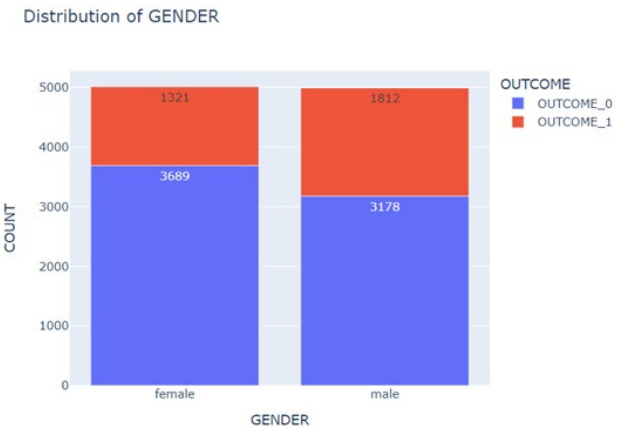
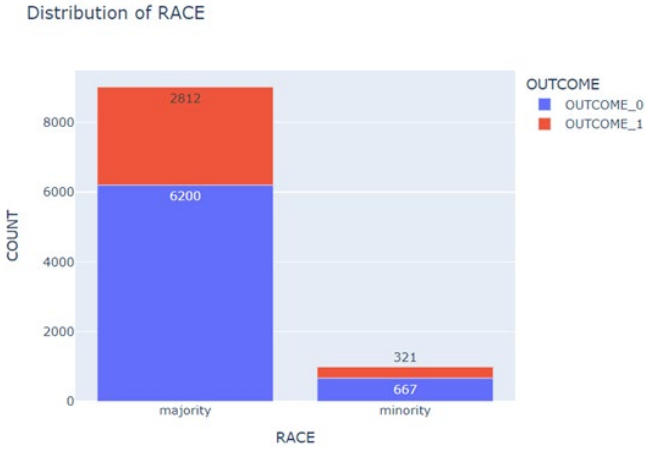


Figure 3 shows the distribution of the gender of the company's policyholders and females exceeds the total number of males by 10. However, the number of approved car insurance claims are significantly higher for males (1812 cases) as compared to females (1321 cases). In other words, 36.31% of males' claim for insurance was successful but only 26.37% females had successful insurance claims.

Figure 4. Race Distribution of Car Insurance Policyholders



Most policyholders of the company are from the majority race (9012 cases) whereas only a small portion of them is from the minority (988 cases) as depicted in Figure 4. However, the proportion of successful insurance claims is slightly higher among the minority, marking at 32.49% as compared to the majority which is at 31.20%.

Figure 5. Driving Experience Distribution of Car Insurance Policyholders

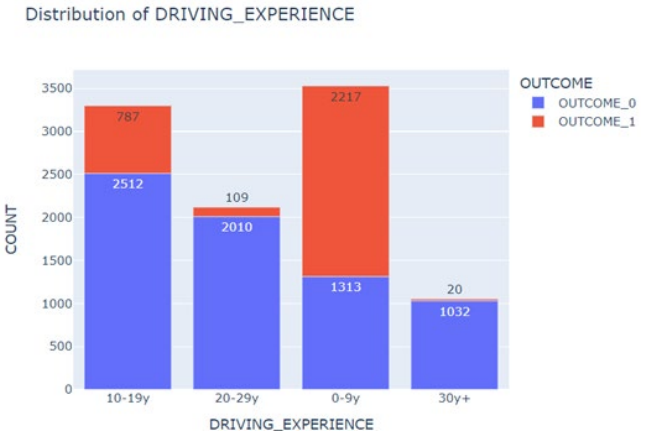
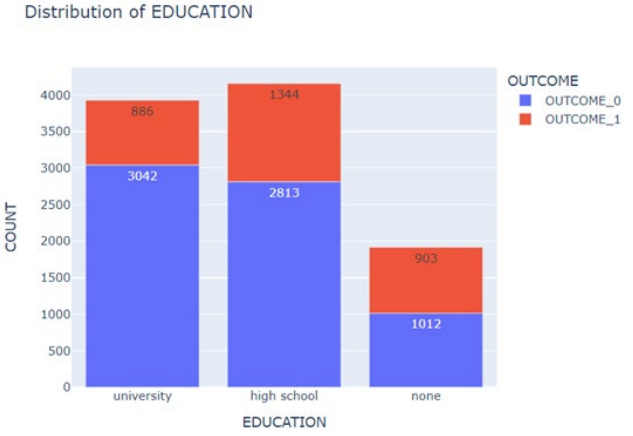


Figure 5 depicts the policyholders' driving experience which is binned into 4 separate groups, and it shows that the majority of the policyholders have a driving experience from 0-9 years (3530 cases) whereas the most uncommon driving experience category among policyholders is 30 years and above. A significant trend is shown where policyholders with less driving experience have a higher success rate to have their insurance claim approved vice versa. The success rate for the category 0-9 years, 10-19 years, 20-29 years, and above 30 years are 62.80%, 23.86%, 5.14%, 1.9% respectively.

Figure 6. Education Type Distribution of Car Insurance Policyholders



The majority of the policyholder's highest education is high school (4157 cases), followed by university (3928 cases) and those who did not receive any education type of education or education that is beyond high school level (1915 cases) as depicted in Figure 6. Nevertheless, most policyholders with education category none received a successful claim of insurance, marking at 47.15%. In

contrast, policyholders with education university and high school had only a 22.56% and 32.33% successful insurance claims respectively.

Figure 7. Income Distribution of Car Insurance Policyholders

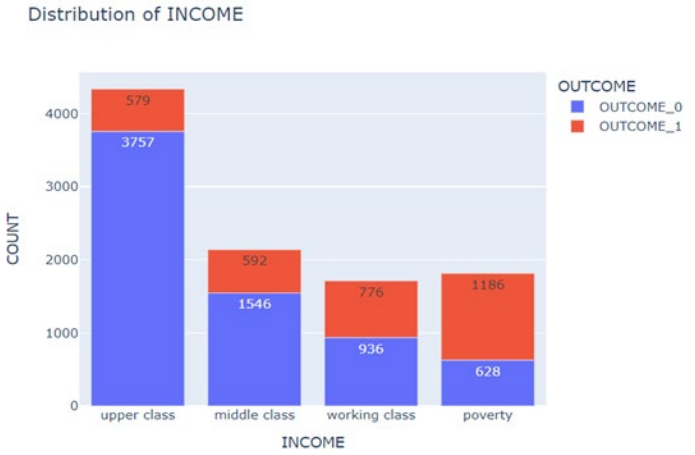
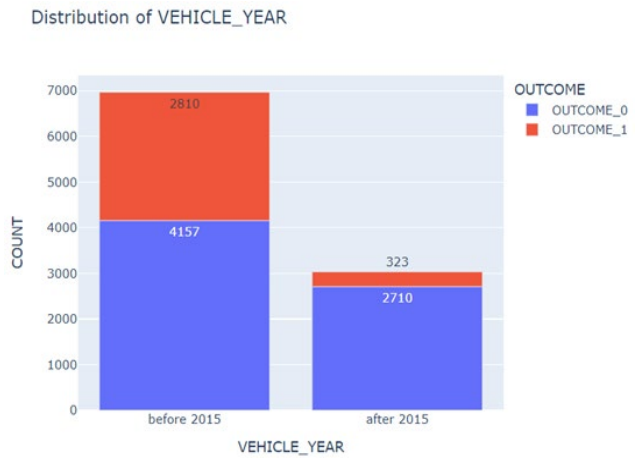


Figure 7 shows the distribution of the income of policyholders and the majority of them are from upper class (4336 cases), followed by middle class (2138 cases), poverty (1814 cases) and working class (1712 cases). However, the correlation between income and the car insurance claim is inversely proportional whereby the percentage of successful car insurance claim by the decreasing sequence of ordinal of income category are 13.35%, 27.70%, 45.33% and 65.38% respectively.

Figure 8. Vehicle Year Distribution of Car Insurance Policyholders



Most vehicle registered under the company's car insurance scheme is purchased before 2015, marking at 6967 cases with high successful car insurance claims (40.33%) as shown in Figure 8. Vehicle that is purchased after 2015 has a lower count, which is a total of 3033 cases and only 10.65% of them were successful claims. Therefore, policyholders with older vehicles are likely to have their claim for insurance to be successful.

Figure 9. Vehicle Type Distribution of Car Insurance Policyholders

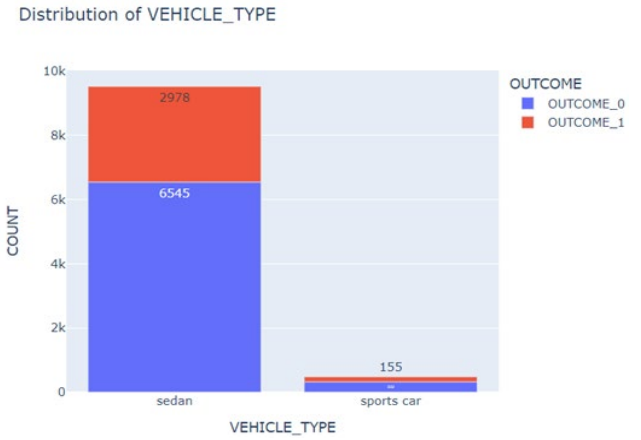


Figure 9 depicts the vehicle type of cars under the company's insurance scheme and most cars are of type sedan, which are 9523 cases out of the 10000 cases. The rest of the 477 vehicles are sports car. However, the type of vehicle with a higher rate of successful insurance claims is sports car, marking at 32.49% as compared to 31.27% in vehicle of type sedan.

Figure 10. Credit Score Distribution of Car Insurance Policyholders

Credit score is one of the continuous attributes in the car insurance dataset with the presence of missing values of 982 as depicted in Figure 10. Credit score is binned into 5 different categories and most of the credit score falls in group 3 (0.0525, 0.235], followed by group 4 (0.235, 0.416], group 2 (0.598, 0.779], group 1 (0.416, 0.598] and group 5 (0.779, 0.961]. Although most of the credit score of policyholders falls at group 3 (4026 cases), but its percentage of successful insurance claims (29.11%) is lower than in group 1 (57.89%) and group 2 (54.19%). Similar incident occurs in group 4 which is the second most common credit score category among policyholders. Despite being the second most common category, its rate of successful insurance claim is lower (16.17%) than those in group 1 and group 2, which are less common among policyholders.

Figure 11. Annual Mileage Distribution of Car Insurance Policyholders

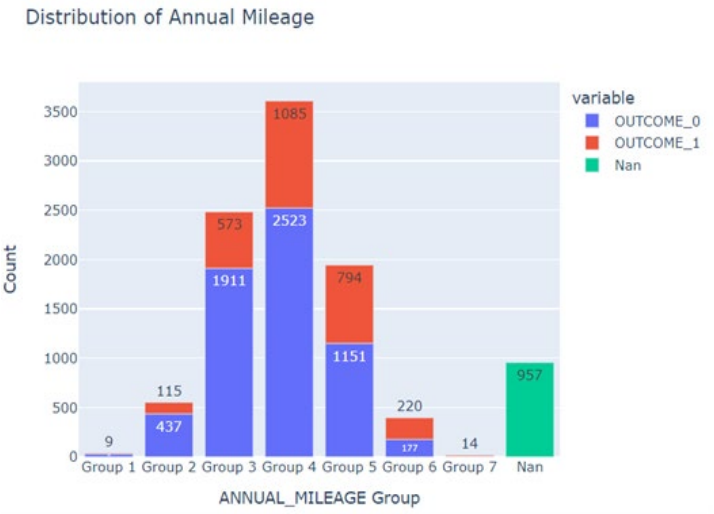


Figure 11 shows the binned distribution of annual mileage of the policyholders' vehicle with the presence of missing value in this attribute. The most common annual mileage category is group 4 (16285.714, 19142.857], followed by group 3 (1980.0, 4857.143], group 5 (13428.571, 16285.714], group 6 (18000.0, 22000.0], group 2 (18000.0, 22000.0], group 6 (10571.429, 13428.571], group 1 (4857.143, 7714.286] and group 7 (7714.286, 10571.429]. A general trend of obtaining a successful car insurance compensation is that a higher annual mileage usage shows a greater chance of successful car insurance compensation with percentage of 23.08%, 20.83%, 23.07%, 30.07%, 40.82%, 55.42%, 77.78%.

Figure 12. Vehicle Ownership Distribution of Car Insurance Policyholders

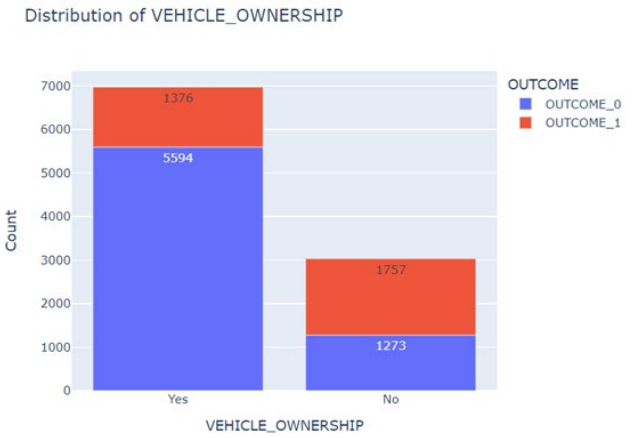
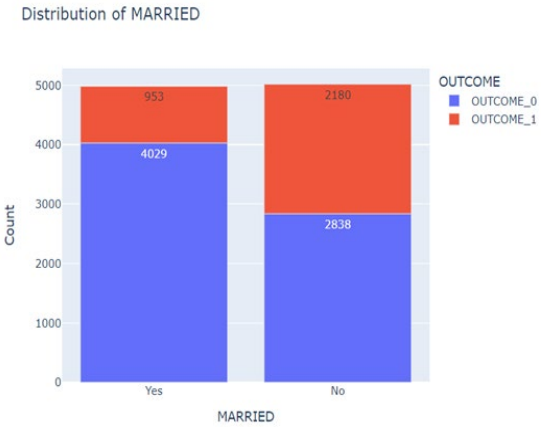


Figure 12 depicts the distribution of vehicle ownership of policyholders and most of them own a car, with 6970 cases while there are only 3030 cases where policyholders do not have a vehicle. Policyholder without ownership to a vehicle is involved in the insurance scheme known as the non-owner car insurance, which is a liability policy for individuals who drives vehicles not of theirs and to keep them insured due to the incurrence of bodily injuries to others, damages to others' vehicle and a legal defence when being sued due to car accidents. Even though most policyholders own at least one vehicle, the chance of getting insured (19.74%) is lower than those who do not own one (57.99%) [12].

Figure 13. Marital Status Distribution of Car Insurance Policyholders



The majority of policyholders are married (4982 cases) while 4218 of them are not as depicted in Figure 13. Nevertheless, only 19.13% of the married policyholders obtained a compensation by the insurance company, which is low when compared to those who are not married (43.44%).

Figure 14. Children Status Distribution of Car Insurance Policyholders

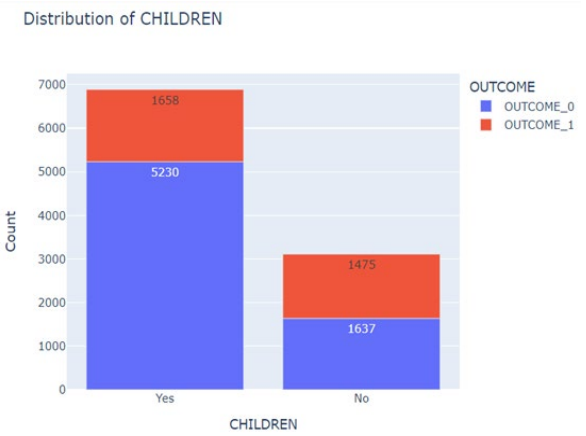


Figure 14 shows records of the presence of children among policyholders or otherwise in relation to the car insurance claim. The majority of policyholders have children (6888 cases) while 3112 out of 10000 of them have no children. However, up to 47.40% of the policyholders with no children were compensated or insured by the company while only 24.07% of policyholders with children were insured.

Figure 15. Postal Code of the Residential Area of Car Insurance Policyholders

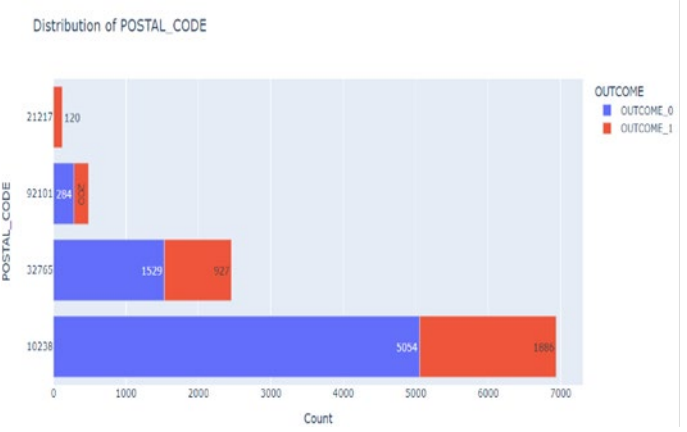


Figure 15 shows 4 different postal codes that reflects on the policyholder's residential area. Most policyholders reside in area with postal code number 10238 (6940 cases), followed by 32765 (2456 cases), 92101 (484 cases) and 21217 (120 cases). The significance in this attribute is all policyholders who lives in the area with postal code 21217 successfully obtained insurance compensation (100%). The next postal code with higher success insurance compensation percentage is 92101 (41.32%), 32765 (41.32%) and 10238 (27.18%).

Figure 16. Number of Speeding Violation breached by Car Insurance Policyholders

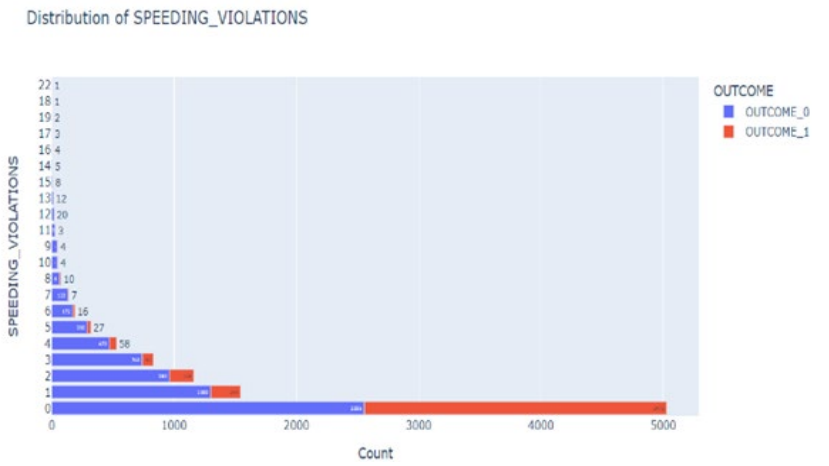
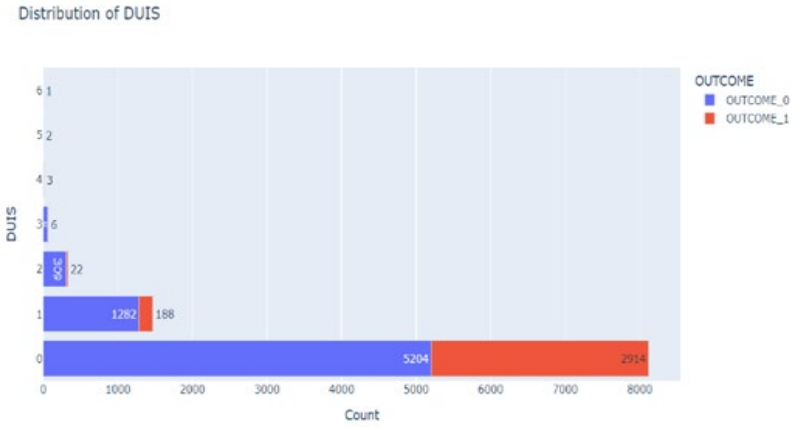


Figure 16 represents the number of speeding violations performed by policyholders, from 0 being the minimum and 22 being the maximum count. The majority of the policyholders was not caught to violate any speed limit (5028 cases) and the total number of policyholders that violated the speeding limit decreases as the count of speeding violation increases. Besides, the correlation between the number of speeding violation and the successful insurance claims is inversely proportional where an increase in the number of speeding violation induces a decrease in the chance of successful insurance claims.

Figure 17. Number of Drinking Under Influence Violation breached by Car Insurance Policyholders



The total number of Driving Under Influence (DUIS) cases violated by policyholders are depicted as in Figure 17. The majority of the policyholders did not drive under influence (8118 cases) and 27.18% of policyholder under 0

DUIS were successfully insured by the company. Generally, the higher the count of DUIS, the lower the chance of policyholders to be successfully insured by the company. However, there is a uniqueness in the data where 30% of policyholders (3 out of 10 people) under DUIS 3 were insured by the company.

Figure 18. Number of Past Accidents experienced by Car Insurance Policyholders

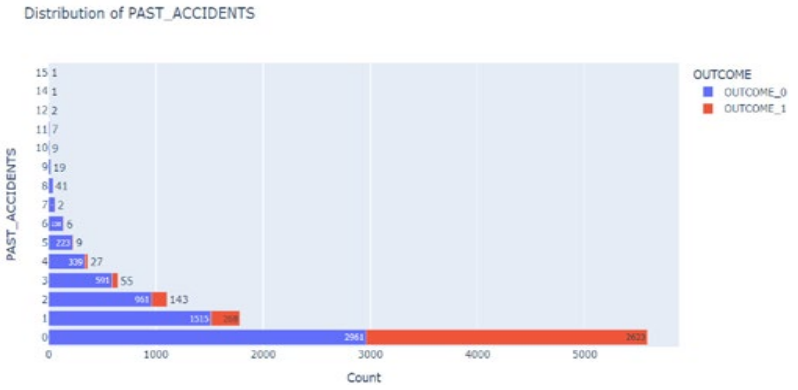


Figure 18 depicts the total number of past accidents experienced by policyholders. The majority of policyholders were not involved in any past accidents (5584 cases) and as the number of past accidents increases, the number of policyholders involved, as well as the chances where policyholders will be insured by the company decreases. Hence, there is a negative correlation between the attribute past accidents and the chances of car insurance claim approval.

3.2 Data Pre-processing

3.2.1 Problems in Car Insurance Dataset

The highlight of the noises in the raw car insurance dataset is all attributes are in their correct and consistent data type and representation without any missing values except for 2 attributes, namely Credit Score and Annual Mileage which has a total missing value of 982 (9.82%) and 957 (9.57%) respectively. Besides, there are attributes with high number of cardinalities such as Annual Mileage, Speeding Violation, Duis, Past Accidents and Credit Score as depicted in a histogram. This leads to the exponential growth of data dimension and its occupation of space and memory, as well as imposes higher risk to limit the performance of a predictive model. Next, there are outliers (unique data) in the attributes of continuous type, specifically Credit Score as shown in Figure 19. In addition, the data in both of the continuous attributes – Credit Score and Annual Mileage are not properly scaled whereby the distances between data points are notably large or different from the rest of the data of other attributes. The

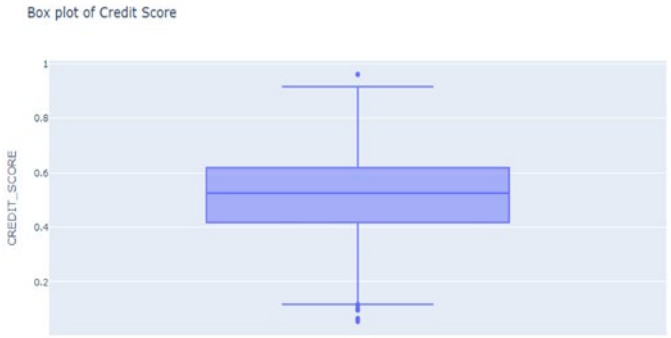
downside is bias may be present in the prediction made by machine learning models, especially in those that considers distances between data points for prediction such as K-Nearest Neighbours (KNN), Support Vector Machine (SVM). These models analyse attributes with a larger range of values to have a stronger role in car insurance approval prediction as compared to those with a smaller range of values. Besides, data of categorical attributes are in a non-numerical form which is a language unintelligible by machine learning models in making an insightful prediction. Next, there is a total of 18 attributes which is applicable in determining the approval of a submitted car insurance claim. However, not every attribute is beneficial and useful, if not will deteriorate the accuracy of the prediction made by the predictive model or the complexity of its interpretation. Last but not least, there is an opportunity to reduce the dimension of data hence reducing its memory and space requirement and complexity, the time needed to reach global minimum in gradient descent (training time), as well as the removal of multicollinearity in the data. Therefore, appropriate data cleaning, transformation and reduction techniques were applied to enhance the quality of data for predictive analysis.

3.2.2 Data Cleaning: Handling missing values, noise and inconsistency in dataset

The 1851 missing values in the raw Car Insurance Dataset, made up from attributes Annual Mileage and Credit Score are handled via Imputation upon transforming categorical and non-numerical data into numerical data via ordinal encoding which will be explained in the data reduction section. Two imputation techniques were tested on the dataset, which are simple imputation and iterative imputation with its measures of central tendency (mean, mode, median). Simple imputation is where all missing values present in an attribute are solely replaced with either of its measures of central tendency whereas iterative imputation modelled an attribute as a function to another and predict the missing values like a regression problem.[13] Since iterative imputation yields a lower mean square error, iterative imputation is used to replace all missing values present in both Annual Mileage and Credit Score attribute with its median.

Noises in the data like outliers as depicted in the Figure 19 boxplot, as well as data in attributes that are highly skewed such as DUIS, past accidents, speeding violations are not removed. The justification for the aforementioned decision is to keep the data as realistic as in real world scenario and enable predictive model to take unique cases into account. Hence, a generalized predictive model to identify if an insurance claim should be approved or otherwise can be developed.

Figure 19. Boxplot of Credit Score Distribution



3.2.3 Data Transformation, Data Scaling, Data Discretization, Feature Selection and Dimensionality Reduction

In data cleaning, categorical and non-numerical attributes are transformed into numerical representation via Ordinal Encoder and such method is chose as such attributes such as Age, Income, Education, Driving Experience have data level of type ordinal. Although the Gender, Race, Vehicle Year, Vehicle Type and Postal Code attributes are not ordinal data but transforming them with Ordinal Encoder works the same to transform them into binary representation or normal numerical representation. With data transformation, it enhances the quality and usability of data for machine learning model training as data is encoded into numbers, which is the sole type of language machines understood.

The cleaned car insurance dataset is then scaled upon imputation with median via Iterative Imputation with Standard Scaler. Data scaling is necessary because there is a significant difference in the scale of values in the column Annual Mileage with other attributes. For instance, the range of Credit Score is (0.053, 0.961) whereas the range of values for Annual Mileage is (2000.0, 22000.0). Such incident as discussed will introduce biases to the predictive model where attributes with a larger range of values plays a stronger role in the car insurance approval prediction as compared to those with smaller range of values. Therefore, data scaling with Standard Scaler overcomes the bias issue by ensuring all continuous numerical values to be at a similar scale, which assists in speeding up the convergence of gradient descent to the local minima.

Next, data discretization is performed with an unsupervised univariate discretization algorithm - K bins discretization on columns Speeding violation, Past accidents, Credit Score and Annual Mileage. Discretization bins continuous variables into k discrete groups, as well as to minimize the dispersion of discrete variables to limit the number of possible states. Hence, this enhances the model prediction and reduces the running time of the developed predictive model.

Data pre-processing phase continues with feature selection to select minimum number of useful attributes during model training, which not only resulted in a better car insurance approval prediction, but it also reduces model training time, complexity and eases model interpretation. Several feature selection techniques were tested on the car insurance dataset, which are the univariate feature selection technique that is dependent on univariate statistical tests to selects k top features according to the k highest scores obtained – SelectKBest then discards unselected or irrelevant attributes, leaving only the k highest scoring features to be used in model training. However, the best value of k that will yield a high-performance predictive model is not known. The next feature selection technique is SelectFromModel which selects the attributes based on the importance of weights defined by the prediction model used. Nevertheless, it is limited to estimators that provides information about features importance attribute and coefficients upon fitting it. The following feature selection technique tested is Sequential Feature Selector that performs forward and backward feature selection from a subset of features in a greedy fashion. Hence, the process of choosing best features involves the continuous addition and removal of features based on cross validation score of the estimator. However, the best value of k that will yield a high-performance predictive model is not known. The last feature selection method tested is the Recursive Feature Elimination with Cross Validation technique that automatically tune the number of features selected with cross-validation. It works by selecting the best subset of features for the specific estimator or predictive model by removing 0 to N features using recursive feature elimination then choosing the subset that yields the best cross validation score. Nevertheless, estimators must have features importance attribute and coefficients attribute after model fitting. Table 5 displays the attributes chosen by each feature selection method and the best feature selection technique is dependent on the predictive model used. In this project, Recursive Feature Elimination with Cross Validation technique which automatically choose the optimum number of attributes used in model training is applied to eliminate insignificant features during model training and ultimately reduces model training time.

Last but not least, 2 different statistical dimensionality reduction techniques namely, Principal Component Analysis (PCA) and Feature Agglomeration were tested on the car insurance dataset. The difference between both methods is PCA focuses on combination of features that captures the variance of the original well that at last creates a low dimensional representation of the dataset. In contrast, feature agglomeration applies hierarchical clustering which builds a dendrogram to pair up features with the highest degree of similarity in their behaviour. While PCA extracts patterns encoding which is based on the highest variance in the dataset, feature agglomeration attempts to separate data into groups and is very sensitive to the measurement of similarity used.[14][15] In

this project, PCA is chose to reduce the dimension of the car insurance data whilst maintaining the principal properties of the attributes of the original data prior predictive model development.

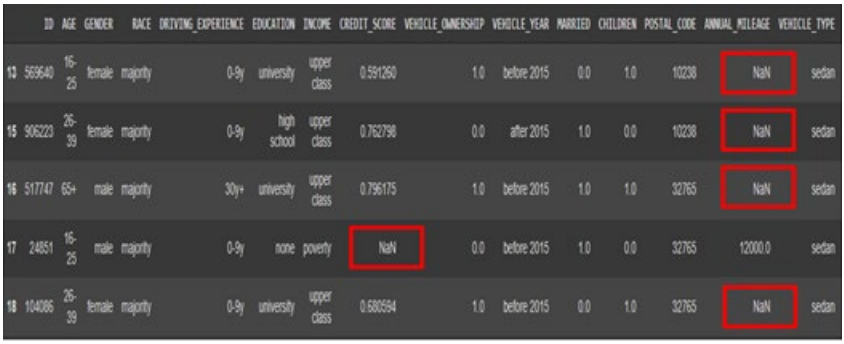
Table 6. Different attributes selected from the main car insurance dataset based on specific feature selection method

Feature Selection Method	Specification	Number of attributes selected	Features Selected
Select K Best	k=10	10	Age, Driving Experience, Vehicle Year, Postal Code, Vehicle Ownership, Married, Children, Credit Score, Speeding Violations, Past Accidents
Select From Model	Model used to retrieve best features: Logistic Regression	5	Gender, Driving Experience, Vehicle Year, Postal Code, Vehicle Ownership
Sequential Feature Selector	k=10, Model used to retrieve best features: K Neighbors Classifier	10	Gender, Driving Experience, Vehicle Year, Vehicle Type, Postal Code, Vehicle Ownership, Married, Annual Mileage,Duis, Past Accidents
Recursive Feature Elimination	Model used to retrieve best features: Support Vector Regressor, Cross validation: 15, number of steps: 1	13	Age, Gender, Race, Driving Experience, Income, Vehicle Year, Postal Code, Vehicle Ownership, Married, Children, Credit Score, Annual Mileage, Past Accidents

4 Results

4.1 Data Cleaning: Handling missing values, noise and inconsistency in dataset

Figure 20. Example of raw data before handling the missing values



	ID	AGE	GENDER	RACE	DRIVING_EXPERIENCE	EDUCATION	INCOME	CREDIT_SCORE	VEHICLE_OWNERSHIP	VEHICLE_YEAR	MARRIED	CHILDREN	POSTAL_CODE	ANNUAL_MILEAGE	VEHICLE_TYPE
13	569640	16-25	female	majority	0-9y	university	upper class	0.591260	1.0	before 2015	0.0	1.0	10238	NaN	sedan
15	906223	26-39	female	majority	0-9y	high school	upper class	0.762798	0.0	after 2015	1.0	0.0	10238	NaN	sedan
16	517747	65+	male	majority	30y+	university	upper class	0.796175	1.0	before 2015	1.0	1.0	32765	NaN	sedan
17	24851	16-25	male	majority	0-9y	none	poverty	NaN	0.0	before 2015	1.0	0.0	32765	12000.0	sedan
18	104086	26-39	female	majority	0-9y	university	upper class	0.680594	1.0	before 2015	0.0	1.0	32765	NaN	sedan

Figure 20 shows the example of raw data before handling the missing values. Both features 'CREDIT_SCORE', and 'ANNUAL MILEAGE' contains missing values by referring to the figure above. The missing values are labelling as 'NaN' (not a number) in the red box.

Figure 21. Example of data after handling the missing values

	AGE	GENDER	RACE	DRIVING_EXPERIENCE	EDUCATION	INCOME	VEHICLE_YEAR	VEHICLE_TYPE	POSTAL_CODE	CREDIT_SCORE	VEHICLE_OWNERSHIP	MARRIED	CHILDREN	ANNUAL_MILEAGE	SPEEDING_VIOLATIONS
13	0.0	0.0	0.0	0.0	2.0	2.0	1.0	0.0	0.0	0.591260	1.0	0.0	1.0	12112.640726	0.0
15	1.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.762798	0.0	1.0	0.0	12125.146787	0.0
16	3.0	1.0	0.0	3.0	2.0	2.0	1.0	0.0	2.0	0.796175	1.0	1.0	1.0	9365.256921	10.0
17	0.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0	2.0	0.394653	0.0	1.0	0.0	12000.000000	0.0
18	1.0	0.0	0.0	0.0	2.0	2.0	1.0	0.0	2.0	0.680594	1.0	0.0	1.0	11211.606770	0.0

Figure 21 shows the example of data after handling the problem of missing values. The technique to impute the missing values is using the multivariate imputation method. The multivariate imputation method is proved that works better in using other features (columns) in the dataset to predict the missing values in the current feature rather than using the non-missing values in a chosen feature and impute the missing values in the same feature using either one statistical measurement (mean/median/mode) of that chosen feature.

Figure 22. Data distribution of each feature (column) in the dataset

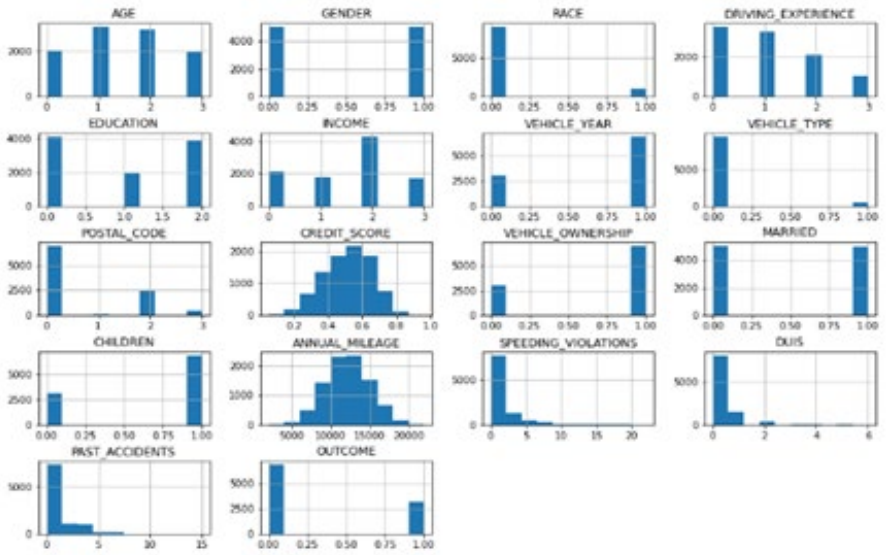


Figure 22 shows the data distribution of each feature in the dataset. Referring to figure 19 and figure 22, attributes such as 'CREDIT_SCORES' consists outliers. Although removing outliers will produce a better-generalized model, it disregards and fails to acknowledge and consider the occurrence of rare/special events such as speeding violations and past accidents that a policyholder may experience. Hence, to develop a machine learning model that is tailored to deal with a real-life situation, the presence of an insignificant number of outliers is not removed.

Figure 23. Example of duplicated data in the dataset

	AGE	GENDER	RACE	DRIVING_EXPERIENCE	EDUCATION	INCOME	VEHICLE_YEAR	VEHICLE_TYPE	POSTAL_CODE	CREDIT_SCORE	VEHICLE_OWNERSHIP	MARRIED	CHILDREN	ANNUAL_MILEAGE
2370	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	NaN	0.0	0.0	0.0	14000.0
3200	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	2.0	NaN	0.0	0.0	1.0	NaN
4457	0.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	NaN	0.0	0.0	0.0	13000.0
5605	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	NaN	0.0	0.0	0.0	NaN
6151	0.0	0.0	0.0	0.0	0.0	3.0	1.0	0.0	0.0	NaN	1.0	0.0	1.0	9000.0
6766	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	NaN	0.0	0.0	0.0	15000.0
7259	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	NaN	0.0	0.0	0.0	14000.0
7400	0.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	NaN	0.0	0.0	0.0	14000.0
7702	0.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	NaN	0.0	0.0	0.0	17000.0
7802	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	NaN	0.0	0.0	0.0	15000.0
8102	1.0	1.0	0.0	1.0	2.0	2.0	1.0	0.0	0.0	NaN	1.0	1.0	1.0	NaN
9150	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	NaN	1.0	0.0	0.0	16000.0

Figure 23 shows the example of duplicated data in the dataset. There are 18 rows of records are classified as duplicated data out of a total of 10,000 rows in the dataset by referring to the figure above. Although there is a number of duplicated data exists in the dataset but each of the duplicated data has its own

unique id. This indicates that each record represents a completely different car insurance policyholder and hence none of the duplicated data will be deleted.

4.2 Data Transformation, Data Scaling, Data Discretization, Feature Selection and Dimensionality Reduction

Figure 24. Example of description of the unstandardized data before data scaling

VEHICLE_YEAR	VEHICLE_TYPE	POSTAL_CODE	CREDIT_SCORE	VEHICLE_OWNERSHIP	MARRIED	CHILDREN	ANNUAL_MILEAGE	SPEEDING_VIOLATIONS	
10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	
0.696700	0.047700	0.648400	0.515373	0.697000	0.498200	0.688800	11693.261990	1.482900	
0.459707	0.213141	1.004828	0.133163	0.459578	0.500022	0.463008	2736.920202	2.241966	
0.000000	0.000000	0.000000	0.053358	0.000000	0.000000	0.000000	2000.000000	0.000000	Min
0.000000	0.000000	0.000000	0.422798	0.000000	0.000000	0.000000	10000.000000	0.000000	
1.000000	0.000000	0.000000	0.523157	1.000000	0.000000	1.000000	12000.000000	0.000000	
1.000000	0.000000	2.000000	0.612793	1.000000	1.000000	1.000000	13815.113333	2.000000	
1.000000	1.000000	3.000000	0.960819	1.000000	1.000000	1.000000	22000.000000	22.000000	Max

Figure 24 shows the example of a description of the unstandardized data before data scaling. The 'ANNUAL_MILEAGE' feature shows a large difference between the range value based on the figure above. The 'ANNUAL_MILEAGE' will dominate over other features during training machine learning model process and it will have more contribution to the computation, just because it has a bigger value compared to other features. So, to prevent this problem, transforming features to comparable scales using standardization is the solution.

Figure 25. Example of description of the standardized data after data scaling

VEHICLE_TYPE	POSTAL_CODE	VEHICLE_OWNERSHIP	MARRIED	CHILDREN	SPEEDING_VIOLATIONS	DUIS	PAST_ACCIDENTS	OUTCOME	CREDIT_SCORE	ANNUAL_MILEAGE	
10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000	1.000000e+04	1.000000e+04	
0.047700	0.648400	0.697000	0.498200	0.688800	1.482900	0.23920	1.056300	0.313300	4.606537e-16	-4.848344e-16	
0.213141	1.004828	0.459578	0.500022	0.463008	2.241966	0.55499	1.652454	0.463858	1.000050e+00	1.000050e+00	
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	-3.469728e+00	-3.541845e+00	Min
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	-6.952346e-01	-6.187052e-01	
0.000000	0.000000	1.000000	0.000000	1.000000	0.000000	0.00000	0.000000	0.000000	5.846004e-02	1.120796e-01	
0.000000	2.000000	1.000000	1.000000	1.000000	2.000000	0.00000	2.000000	1.000000	7.316242e-01	7.753085e-01	
1.000000	3.000000	1.000000	1.000000	1.000000	22.000000	6.00000	15.000000	1.000000	3.345284e+00	3.766004e+00	Max

Figure 25 shows the example of a description of the standardized data after data scaling. The 'ANNUAL_MILEAGE' feature is scaled by removing the mean and scaling to unit variance. Hence the features can dominate equally to the model and the computation time of the training decreases and also it does yield better results.

Figure 26. Data distribution of each feature (column) in the dataset after data discretization

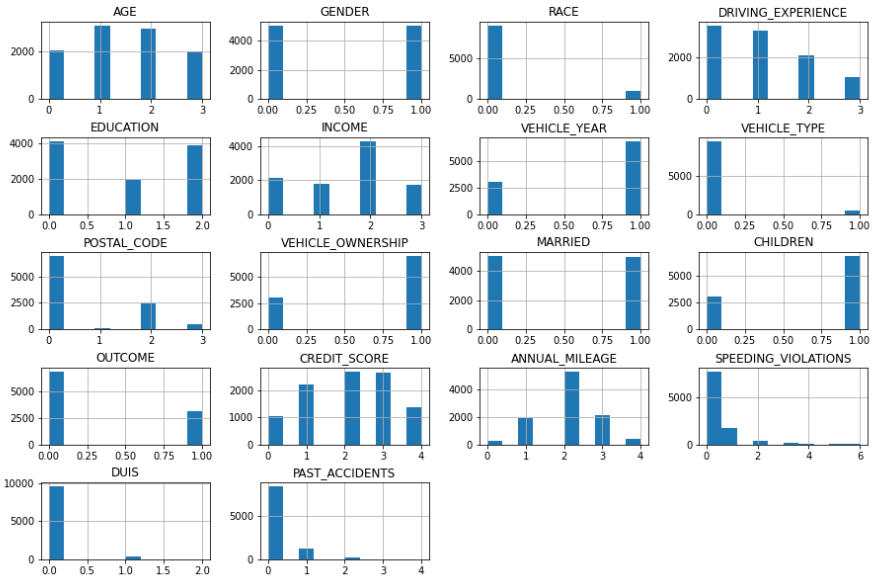


Figure 26 shows the data distribution of each feature in the dataset after data discretization. Numerical input variables may have a highly skewed or non-standard distribution. This could be caused by outliers in the data, multi-modal distributions, highly exponential distributions, and more. Many machine learning algorithms prefer or perform better when numerical input variables have a standard probability distribution. The discretization transform provides an automatic way to change a numeric input variable to have a different data distribution, which in turn can be used as input to a predictive model. These features such as 'CREDIT_SCORE', 'ANNUAL_MILEAGE', 'SPEEDING_VIOLATIONS', 'DUIS' and 'PAST_ACCIDENTS' are continuous data, and each feature has different distribution compared to one another by referring to the figure 22. Therefore, these features are going to be discretized to achieve a standard distribution.

Figure 27. Example of selected features after feature selection technique

	AGE	GENDER	RACE	DRIVING_EXPERIENCE	INCOME	VEHICLE_YEAR	POSTAL_CODE	VEHICLE_OWNERSHIP	MARRIED	CHILDREN	CREDIT_SCORE	ANNUAL_MILEAGE	PAST_ACCIDENTS
0	3.0	0.0	0.0	0.0	2.0	0.0	0.0	1.0	0.0	1.0	3.0	2.0	0.0
1	0.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	3.0	0.0
2	0.0	0.0	0.0	0.0	3.0	1.0	0.0	1.0	0.0	0.0	2.0	2.0	0.0
3	0.0	1.0	0.0	0.0	3.0	1.0	2.0	1.0	0.0	1.0	0.0	2.0	0.0
4	1.0	1.0	0.0	1.0	3.0	1.0	2.0	1.0	0.0	0.0	1.0	2.0	0.0

Figure 27 shows the example of selected features after feature selection technique. The data features used to train machine learning models have a significant impact on the performance of the model. Model performance can be harmed by features that are irrelevant or only partially relevant. As a result, reducing the number of input variables is desirable to both reduce the computational cost of modelling and, in some cases, improve the performance of the model. A total 13 out of 18 features are selected by using the recursive feature selection with cross validation method. These features are determined as the most relevant features that have the strongest relationship with the target variable and able to contribute much to the model hence the model can generate an accurate outcome.

Figure 28. Example of data after dimensionality reduction

	AGE	GENDER	RACE	DRIVING_EXPERIENCE	INCOME	VEHICLE_YEAR	POSTAL_CODE	VEHICLE_OWNERSHIP	MARRIED	CHILDREN	CREDIT_SCORE	ANNUAL_MILEAGE	PAST_ACCIDENTS
0	0.953343	-0.622033	-0.354414	-0.898978	-0.384870	1.594129	0.184547	0.223370	0.500337	-0.701112	0.532796	0.078642	-0.136456
1	-2.537095	-0.802031	0.146777	-0.147050	0.303880	-0.340219	0.361956	-0.381019	0.249873	0.249644	0.022340	-0.279917	-0.108320
2	-1.386940	-0.290501	-1.482803	-1.232866	-0.422788	-0.479020	-0.578699	0.117883	-0.391011	-0.245258	0.527691	-0.291033	-0.089770
3	-2.382989	1.825815	-1.578011	0.208343	0.042549	-0.229941	0.406163	0.232866	-0.369905	-0.603366	-0.337437	0.256680	-0.116753
4	-0.886031	1.549266	-1.416426	0.222215	0.687438	-0.327309	0.337817	0.076373	-0.445377	-0.195532	0.264539	-0.558849	-0.091635

Figure 28 shows the example of data after dimensionality reduction. The dimensionality reduction of the data is done by principal component analysis (PCA) technique. Principal component analysis (PCA) simplifies the complexity in high-dimensional data while retaining trends and patterns. It does this by transforming the data into fewer dimensions, which act as summaries of features.

5 Conclusion

The domain of this research is regarding the prediction of the approval of car insurance claims, which is significant in protecting our cars during accidents. It also protects us from financial liability, medical expenses as well as legal consequences. In this research, our focus is to categorize the customer according to their attributes to decide if they are able to claim their car insurance. In addition, this research discusses various solutions and advice given by different authors for problems encountered in similar domains.

A real-world car insurance dataset is downloaded from Kaggle website, and the data will be used to train a classifier machine learning model upon data pre-processing stage. Data pre-processing starts with Exploratory data analysis (EDA), which is an important step to detect any data anomalies, as well as to discover and understand different patterns in the data. In this research, the general information and characteristics of the data is extracted during EDA and is well presented and discussed in data descriptive analysis and data quality reports. With EDA, a better understand on the suitable data pre-processing techniques to apply to the car insurance data set is achieved with minimum error.

From EDA, it is discovered that the car insurance dataset contains 18 features/attributes and 1 target attribute for each of the 10000 customers of the insurance company. Unfortunately, there are several problems in the dataset, namely the presence of missing values in some records, outliers, appear in unsuitable format for model training, high number of cardinalities and unnormalized distribution of data. Hence, this justifies the need for data pre-processing that converts unclean raw data into a usable, comprehensible format. With a complete, clean and high-quality data, it eases machine learning models to discover hidden insights in the data and enhance the efficiency of model training. Therefore, this research paper has discussed various data pre-processing techniques to each problem in the data set. This research paper also conducts a direct comparison of these techniques to discover the best possible technique to clean the raw data. Lastly, the result of data pre-processing is shown in this research paper.

6 Acknowledgement

The authors would like to thank Dr. Azuraliza Abu Bakar for her advice, patience and guidance throughout the work of this research. The code implemented can be found in
https://colab.research.google.com/drive/1eVTMPA7Y5pbF8Bxff587NTPSFMiG34cP?usp=sharing

Tang Jia Hui is a computer science undergraduate from the National University of Malaysia (UKM), specializes in data science studies. Her contribution to the research focuses on research methodology and coding of this study of car insurance approval prediction.
Contact: a176297@siswa.ukm.edu.my

Lim Ka Li is a computer science undergraduate from the National University of Malaysia (UKM), specializes in data science studies. Her contribution to the research focuses on the abstract, introduction, literature review of car insurance approval prediction.
Contact: a176496@siswa.ukm.edu.my

Chong Yu Jie is a computer science undergraduate from the National University of Malaysia (UKM), specializes in data science studies. His contribution to the research focuses on the result of presentation and visualization after the data pre-processing and conclusion.
Contact: a176730@siswa.ukm.edu.my

7 References

[1] Kousky, C. (2013). Revised Risk Assessments and the Insurance Industry. Insurance Law.

[2] Auto Insurance Guide. https://www.coverage.com/insurance/auto/guide/

[3] Sarker, I.H. Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. SN COMPUT. SCI. 2, 377 (2021). https://doi.org/10.1007/s42979-021-00765-8

[4] Theodoros Iliou, Christos-Nikolaos Anagnostopoulos, Marina Nerantzaki, and George Anastassopoulos. 2015. A Novel Machine Learning Data Preprocessing Method for Enhancing Classification Algorithms Performance. In Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS) (EANN '15). Association for Computing Machinery, New York, NY, USA, Article 11, 1–5. https://doi.org/10.1145/2797143.2797155

[5] Hanafy, Mohamed, and Ruixing Ming. 2021. Machine Learning Approaches for Auto Insurance Big Data. Risks 9: 42. https://doi.org/10.3390/risks9020042

6. Abdelhadi, S., ElBahnasy, K.A., & Abdelsalam, M.M. (2020). A PROPOSED MODEL TO PREDICT AUTO INSURANCE CLAIMS USING MACHINE LEARNING TECHNIQUES.

7. Endalew Alamir, Teklu Urgessa, Ashebir Hunegnaw and Tiruveedula Gopikrishna, “Motor Insurance Claim Status Prediction using Machine Learning Techniques” International Journal of Advanced Computer Science and Applications (IJACSA), 12(3), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0120354

8. Dan Huangfu. April 2015. Data Mining for Car Insurance Claims Prediction. A Project Report for Worcester Polytechnic Institute

9. Seema Rawat, Aakankshu Rawat, Deepak Kumar, A. Sai Sabitha, Application of machine learning and data visualization techniques for decision support in the insurance sector, International Journal of Information Management Data Insights, Volume 1, Issue 2, 2021, 100012, ISSN 2667-0968, https://doi.org/10.1016/j.jjimei.2021.100012.

(https://www.sciencedirect.com/science/article/pii/S2667096821000057)

10. Tim Pijl. August 2017. A Franework to Forecast Insurance Claim. A thesis for the Erasmus University Rotterdam

11. Hui Dong Wang, Research on the Features of Car Insurance Data Based on Machine Learning, Procedia Computer Science, Volume 166, 2020, Pages 582-587, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.02.016. (https://www.sciencedirect.com/science/article/pii/S1877050920301381)

12. Metz, J. (2022, April 5). Non-owner car insurance: Do you need it? Forbes. Retrieved May 1, 2022, from https://www.forbes.com/advisor/car-insurance/non-owner/

13. Brownlee, J. (2020, August 18). Iterative imputation for missing values in machine learning. Machine Learning Mastery. Retrieved May 1, 2022, from https://machinelearningmastery.com/iterative-imputation-for-missing-values-in-machine-learning/Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Mor- gan Kaufmann, San Francisco (1999)

14. F. Pedregosa, G. Varoquaux, A. Gramfort, V. T. B. Michel, O. Grisel, M. Blondel, P. W. R. Prettenhofer, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, Scikit-learn: Machine Learning in Python, vol. w12, Journal of Machine Learning Research, 2011, pp. 2825-2830.

15. (2016, February). A comparison between PCA and hierarchical clustering. KDnuggets. Retrieved May 1, 2022, from https://www.kdnuggets.com/2016/02/qlucore-comparison-pca-hierarchical-clustering.html