Predictive Model of Future Insurance claims By using Machine Learning Algorithms

Muhammad Atta Khan (19k-0932)
National University of Computer and Emerging Sciences
Karachi, Pakistan
Email:19k0932@nu.edu.pk
Abbas Ali Rizvi (18k-1312)
National University of Computer and Emerging Sciences
Karachi, Pakistan
Email:18k1312@nu.edu.pk

I. Research Goal

The aim of this research is to apply machine learning algorithms on the data set of Insurance to identify relation between claim and insurance premium data and to visualize the outcomes using different visualization techniques. The goal is to identify hidden patterns from the raw data and then predict a best fit model to predict future values of data sets . The data will be classify into different classes and then test again different data sets.

II. Introduction

Data Science is the area of study which involves extracting insights from vast amounts of data by the use of various scientific methods, algorithms, and processes. It helps you to discover hidden patterns from the raw data.

The term Data Science has emerged because of the evolution of mathematical statistics, data analysis, and big data. Data Science is an interdisciplinary field that allows you to extract knowledge from structured or unstructured data. Data science enables you to translate a business problem into a research project and then translate it back into a practical solution. Key components of data science are given in figure 1.

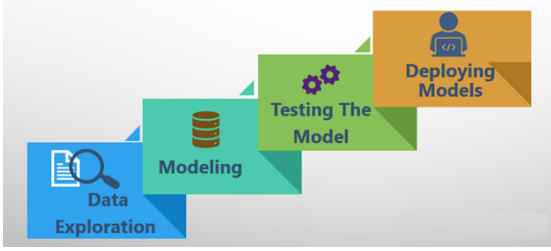


Figure 1. Data science key components

Data science process is consists of six key modules discovery, data preparation, model planning, model building, operation and communicate results. These processes are given in figure 2.

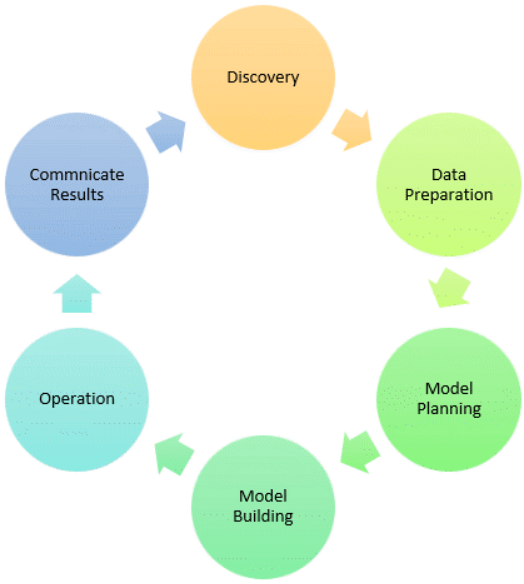


Figure 1. Data science process model

A. Discovery

Discovery step involves acquiring data from all the identified internal and external sources which helps you to answer the business question.

The data can be:

- Logs from webservers.
- Data gathered from social media.
- Census datasets.
- Data streamed from online sources using APIs.

B. Data Preparation

Data can have lots of inconsistencies like missing value, blank columns, incorrect data format which needs to be cleaned. You need to process, explore, and condition data before modeling. The cleaner your data, the better are your predictions.

C. Model Planning

In this stage, you need to determine the method and technique to draw the relation between input variables. Planning for a model is performed by using different statistical formulas and visualization tools. SQL analysis services, R, and SAS/access are some of the tools used for this purpose.

D. Model Building

In this step, the actual model building process starts. Here, Data scientist distributes datasets for training and testing. Techniques like association, classification, and clustering are applied to the training data set. The model once prepared is tested against the ”testing” dataset.

E. Operationalize

In this stage, you deliver the final baselined model with reports, code, and technical documents. Model is deployed into a real-time production environment after thorough testing.

F. Communicate Results

In this stage, the key findings are communicated to all stakeholders. This helps you to decide if the results of the project are a success or a failure based on the inputs from the model.

III. Data Collections

Data is collected from the reputed insurance company in the form of excel sheets. Data is in its raw form and divided into separate file for claims and premiums. The number of rows in data set is around 6 million and consist of more than 120 different features.

IV. Data Preparation

The data is in raw form and have lots of inconsistency in the form of missing value, blank values, incorrect format of data values. The cleaning process is starts of finding the missing values and blanks fields. The majority of features columns contain missing values but their ration is quite low and around 2 to 3% of total data is missed. The features of data are mostly categorical so we decided to drop the missing values or blank values from the data set. The same procedure is follow for most of the cases of incorrect values but in few cases the incorrect values is replaced by abbreviations. By following these steps data got cleaned from all inconsistencies but still it is not ready for use as a source model data.

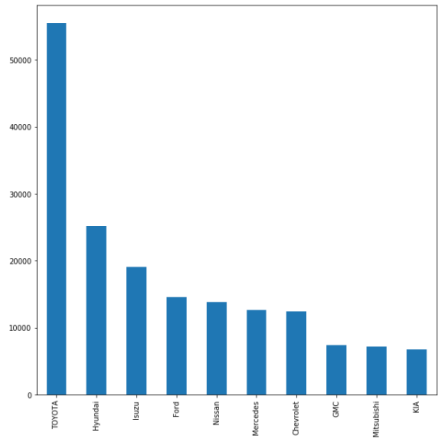
As most of the features set of data is categorical we need to first convert them into either numerical or in binary form. The categorical feature set is converted into binary form using one hot encoding. One hot encoding ensure equal weights for each class of data and convert the classes into binary form. It convert each classes into virtual feature and add them as column as data columns. After cleaning and preparation the data is converted into the form that could be provided to the machine learning algorithms for the prediction.

V. Data Exploration

After cleaning of data different key features is identified and relationship between those feature sets is then observed using different visualization methods. Some of the key relationship are as follows.

A. Car Companies ranking by claim frequency

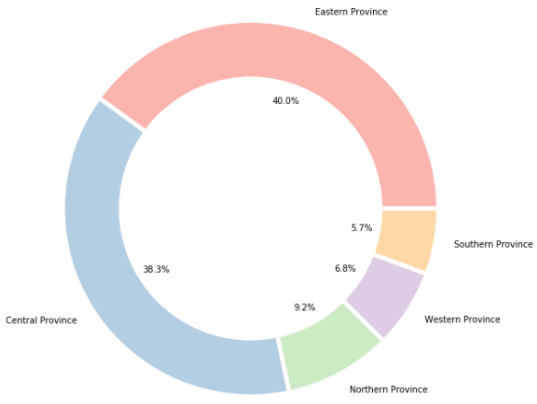
One of the key data set feature is to group down the frequency of claims occurring with respect to the car companies models. The visualization of this relationship required car companies on x-axis and frequency count on y-axis. It is observed that the claim ratio is quite high with Toyota company cars and there is a big gap between the first and the second company. After that the ratio decreases gradually along all the companies.



car models vs frequency

B. Claims frequency with respect to region

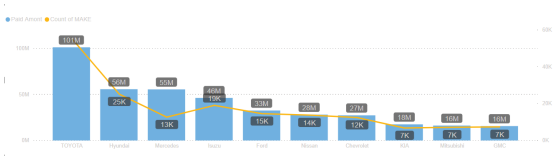
Another insight in data is to visualize the frequency of claims with respect to the regions. The analyses identify the frequency in percentage in particular regions. It was analyze that around 78% of overall claims are register from Eastern and Southern region. The analyses is then visualize using circle or pie graphs



frequency vs regions

C. Paid Amount and count of MAKE with respect to MAKE

Total paid amount of claims and the count of the car model in overall data is another insight from the data set. The visualization is the combination of line and bar and graph in which Amount paid is observed by the bars of graph and the count of MAKE is displayed as a line over it. It is observed that the Toyota company has the highest number of claims recorded and they get paid over 101M as claims.



Paid Amount and count of MAKE vs MAKE

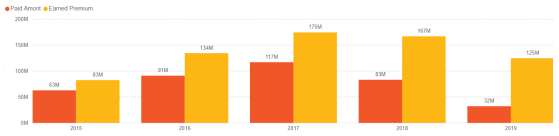
D. Chain Ladder of Paid Claims

The Chain Ladder Method (CLM) is a method for calculating the claims reserve requirement in an insurance company's financial statement. CLM is used to construct the triangle of paid claim between different years and development year lags. The values of chain ladder triangle is the amount paid in each development lag.

Years	0	1	2	3	4
2015	48,573,998	17,692,209	2,059,100	796,006	34937
2016	43,798,700	14,986,929	4,705,531	658,851	
2017	111,554,369	20,593,599	786,101		
2018	64,527,696	14,285,229			
2019	69,655,152				

E. Paid Amount and Earned Premium by Year

The comparison between the Earned premium and the paid claim is the most important insight and it shows the amount of profit and loss of a company. The data available for this analyses is in between year 2015 – 2019. It is observed that the premium amount collected by company throughout the year is higher than the paid claim amount.



Paid Amount and Earned Premium vs Year

VI. Data Modeling

The data cleaning and exploration process lead to the data modelling . In this process you need to provide clean and transform data to your model to predict the best results as outcome . In this research we have performed two types of modelling.

A. Paid Claims Future Value Predictor

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression model is based on the input variables. Simple linear regression is used when there is one independent variable and one dependent variable. The linear equation is constructed using intercept and slope values.

$$Y = a + bX$$

If there is more than one independent variable then multivariate linear regression is applied. The linear equation is the sum of intercept and the product of all independent variables slope and independent variable value.

$$Y = a + b_1X_1 + b_2X_2 + \cdots$$

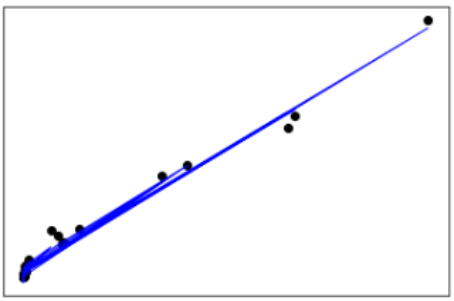
As there are multiple variable identified for model prediction so we have decided to apply multiple linear regression on our data set. The idea is to predict the cumulative sum of paid claims . The outcome is predicted against the numbers of independent input variables which includes reported year , development lag , frequency of claims in each development lag and year and the development year. The data is divided into training and testing sets . additional sample data is also constructed to predict the future years and development lag paid claims values . The table below show the comparison actual and predicted value and the difference between both values in percentage.

Years	devlag	actual values	predicted values	error %
2015	0	485M	501M	3.09 %
2015	1	145M	152M	4.82%
2015	2	65M	69M	5.7%
2015	3	18M	19M	4.2%
2015	4	7M	7.4M	4.1%
2019	0	721M	735M	1.80%
2019	1	400M	410M	2.5 %

The root mean square error (RMSE) is around 500 whereas square root of RMSE is 0.98 . The slope and intercept are as follows .

$$Slope : [[0.02094718 - 14.44891118 - 7.95469908]]$$

$$Intercept : [16092.68844961]$$



frequency vs predicted values

B. Classification on Claim types

Random forest is a Supervised Learning algorithm which uses ensemble learning method for classification and regression. Random forest is a bagging technique and not a boosting technique. Random forest create forest of trees and then classify on the bases of provided sample sets. Random forest avoid over fitting and increase the predictive power of algorithm.

For providing data for model we have first convert all the categorical data in binary with the help of one hot encoding technique and the data is then provided to the model. The number of trees for forest is set to 100 with 1000 samples of training data. The data set is divided into training and testing with the ratio of 70% and 30 % respectively. After setting up the model the test data is then provided and the accuracy of model is calculated. It is observed that the accuracy of the model is 60% approximately.

VII. Results and Conclusion

It is concluded that the pattern of paid Amount of claims is decreasing with respect to the development lag and the development year. The classification model is able to predict around 60 % data correctly. The different exploratory visualization is suggesting the the maximum amount of premium collected is used in paying the claim registered. The company need to review its premiums deals and try to collect more premium for profit.