Risk Classification Using Supervised Learning Techniques

Brendan Pham @ UMN - Twin Cities | brendanbpham@gmail.com | April 14,2023

Abstract: Insurance is a product sold by insurance companies, which insures the customers stability by transferring loss risk over to the company. In insurance, risk classification involves accessing the likelihood a policyholder will file a claim. The challenge of the insurance companies is the arbitrary nature of events that cause loss, and many times these losses cost a considerable amount to the insurance companies. The fundamental goal of this project can be summarized into:

- What are some techniques to evaluating and predicting risks
  - Kangaroo Insurance Auto Insurance Modeling Problem
- How effective are these models in evaluating and predicting risks

Keywords: risk classification; regression model; tweedie distribution

---

1.1 Background & Context

Traditionally, the role of underwriters within health, and property & casualty insurance companies worked closely with pricing actuaries to assess and manage risks associated with covering their clients. Their task was to determine a pure premium at which an agreed level of risk coverage can be provided by the insurance company per risk management policy. Historically, underwriters and actuaries would have to manually review information to decide price per ratemaking for a policy. To determine risk they would have determine:

- what type of risk the insurance company agrees to insure,
- What rate can be charged to policyholder that covers for any level of risk

Presently, the underwriting and actuary process is becoming automated, supported by “machine and deep learning models built within the companies technology structure” (Hubbel) The result of the influx of data science being introduced to insurance companies has allowed faster data-driven decisions regarding pricing. Insurance companies are looking for new data to determine client specific risks, for example utilizing trackers for car insurance has made a large impact on determining pricing.

1.2 General Linear Models & General Additive Models

General Linear Models: Insurance companies have been utilizing the General Linear Models for risk classification. General Linear Models have become more utilized thanks to its powerful predicting tool for insurance companies, thanks to its first introduction when used by the insurance company Progressive to model credit risk.

The General linear model is a supervised machine algorithm that utilizes the relationship between a response variable (classes,binary, continuous) we want to predict given explanatory variables. A Generalized linear model consists of three components: random components, systematic components, and a link function. A random component is the probability distribution of the target variable. The most common probability distributions are normal distributions, binomial, poisson, and binary. The systematic component revolves around the explanatory variables where each can be continuous, discrete, or even classes. The link function transforms the response variable to a nonlinear relationship.

General Linear models have become a powerful tool for risk classification for insurance companies because of the models' flexibility, customization, efficiency, accuracy, and interpretability. Hubbel has stated that "often General linear models's outperform other commonly used statistical models." In the case of the project, in property & casualty, general linear models have become an effective predicting tool in targeting claim Frequency (claims per exposure), claim severity (dollars of loss per claim or occurrence), and pure premium (dollars of loss per exposure)

The generalized linear model framework allows changes in inputs to affect the output variable:

$$\widehat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

- y = Dependent variable
- β = Beta Weights
- x = regressor
- ϵ = residual

- Common Response distributions: normal, binomial, poisson, and gamma, and gaussian.

- Response variable: in GLMs the predicted response variables can be continuous, count, binary
- Predictor Variables: Can be continuous, categorical, and interactions
- Link function: mean of the response variable dependent on the predictor. Often a function to map the predictor to the mean of the response variable

General Additive Models: Insurance companies have been utilizing the General Linear Models for risk classification. General Additive Models have become more utilized thanks to its powerful predicting tool for insurance companies when predicting risk for policies. Insurance companies have utilized general additive models for risk classification thanks to its flexibility in dealing with non-linear relationships.

The General Additive model is a supervised machine algorithm that utilizes the relationship between a response variable (classes,binary, continuous) we want to predict given explanatory variables. A Generalized linear model consists of three components: random components, systematic components, and a link function. A random component is the probability distribution of the target variable. The most common probability distributions are normal distributions, binomial, poisson, and binary. The systematic component revolves around the explanatory variables where each can be continuous, discrete, or even classes. The link function transforms the response variable to a nonlinear relationship. This is seemingly similar to the general linear model, but additive of functions to the explanatory variables. Smoothing spline functions and polynomial functions help capture complex relationships between the response and explanatory variables that are non-linear.

General additive models have become a powerful predictive modeling technique in risk classification because of its flexibility, non-parametric (can violate assumptions) , and interoperability. Compared to general additive models it is often less efficient, but more accurate since GAMs capture nonlinear relationships between independent and dependent variables which improves the fitting/accuracy of the model while utilizing general linear model properties. Overall general additive models is much more powerful tool for risk classification because of its flexibility in it’s non-linear approach

The generalized additive model framework allows functions that reflect on the relationship of outputs:

$$\widehat{y} = \beta_0 + f(x_1) + f(x_2) + f(x_3) + \epsilon$$

- y = Dependent variable
- β = Beta Weights
- x = regressor
- ϵ = residual
- $f()$ = functions of piecewise, splines, etc

- Response distributions: normal, binomial, poisson, and gamma

- Response variable: in GLMs the predicted response variables can be continuous, count, binary
- Predictor Variables: Can be continuous, categorical, and interactions
- Link function: mean of the response variable dependent on the predictor. Often a function to map the predictor to the mean of the response variable

2.0 Kangaroo Insurance Company Modeling Problem

The dataset is a stylized dataset from travelers. You represent Kangaroo Auto Insurance Company, an Australian company. Your business partner, who is not familiar with statistics at all, would like you to create a rating plan based on the historical auto claim data. Your business partner is concerned about segmentation as well as competitiveness, as there are several other competitors in the market. For this case competition, your group's task is to provide a method for predicting the claim cost for each policy and to convince your business partner that your predictions will work well.

2.1 Objective

Auto Insurance must be priced effectively to reflect on policyholders' underlying level of risk if they were to receive that amount. Test different models in effectively predicting the frequency, severity, pure premium.

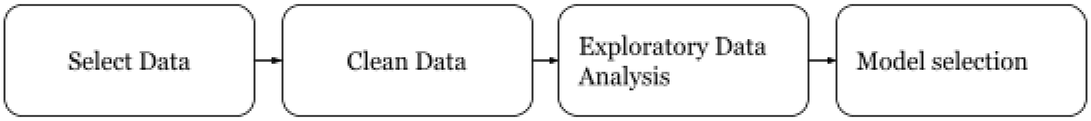
$$\text{Claim Frequency} = \frac{claim\ count}{exposure}$$

$$\text{Claim Severity} = \frac{claim\ cost}{claim\ count}$$

$$\text{Pure Premium} = \quad claim\ frequency \ * \ claim\ severit$$

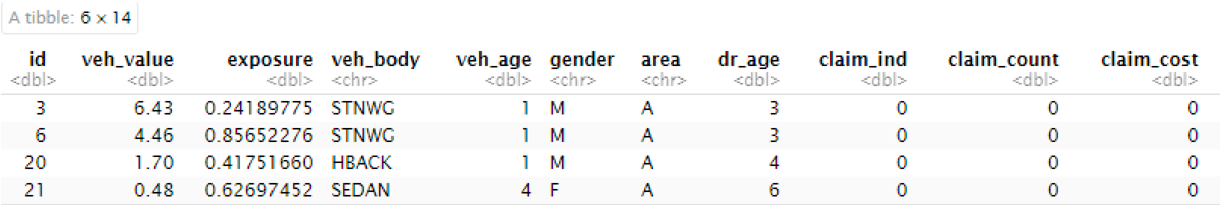
The steps below were taken in order to come to our final model selection, and exploration of the kangaroo dataset.

Figure 1. Data Modeling Process



2.2 Data Selection

Figure 2. Data Table



A tibble: 6 x 14

id	veh_value	exposure	veh_body	veh_age	gender	area	dr_age	claim_ind	claim_count	claim_cost
<dbl>	<dbl>	<dbl>	<chr>	<dbl>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
3	6.43	0.24189775	STNWG	1	M	A	3	0	0	0
6	4.46	0.85652276	STNWG	1	M	A	3	0	0	0
20	1.70	0.41751660	HBACK	1	M	A	4	0	0	0
21	0.48	0.62697452	SEDAN	4	F	A	6	0	0	0

Data Description: The dataset was provided by Travelers case competition. The Kangaroo data set is based on one-year vehicle insurance policies from 2004 to 2005. Data was split into training and validation sets.

Variable Information on the table

- **ID:** policy key
- **veh_value:** market value of the vehicle by $10,000
- **Exposure:** the unit of risk inherited by an insurance premium
- **Veh_body:** type of vehicle
- **Gender:** gender of driver
- **Area:** location of driver
- **dr_age:** Drivers age 1-6, (1) being youngest while (6) is oldest
- **claim_ind:** indicator of claim (0 = no, 1 = yes)
- **claim_count :** the number of claims

- **Claim_cost:** amount of claim that was file

2.2 Data Cleansing & Feature Engineering

As a first step, data was selected, where the following information was cleansed by the following steps:

- Data Standardization: standardizing formats of data, ensuring the consistency and homogenous data. Example changing claim_cost and veh_value to numerator, and exposure to a double.
- Searching for missing values, blank spaces, Non-applicable, or null. As far as missing data there was missing data relevant to claim costs.
- Filtering in order of policy of ID and detecting repeated instances of a policy ID. The filter found no instances in this problem, since the dataset was constructed carefully.
- Outliers: detecting outliers of existing dataset: only known issue came to claim count

2.3 Model selection for risk classification

The following models were selected to predict claim frequency, claim costs, claim severity, and loss cost

Logistic Regression: Logistic Regressions a generalized form of the standard logistic regression model. Given an instance X, input features are supplied to the model in order to compute a score for each class - then, the probability of belonging to each class is estimated by applying the link function. Then, the algorithm returns a prediction as the class with the highest estimated probability.

GAM: In statistics, a generalized additive model (GLM) is a flexible generalization of ordinary linear regression. General additive models like a linear model, but apply linear combinations of non-linear functions of the individual features. Useful when the response to some features are nonlinear. In this project I will be utilizing spline functions

Random Forest Classifier: Supervised machine learning algorithm that is used for this purpse of the project to predict two cases utilizing ensemble learning methods. Random forest trains multiple decision trees, and random sampling with replacement. The decision trees are averaged across all predictions to produce a final prediction. Averaging models has the benefit of reducing overfitting. RAndom forest has become a staple machine learning model for data scientists.

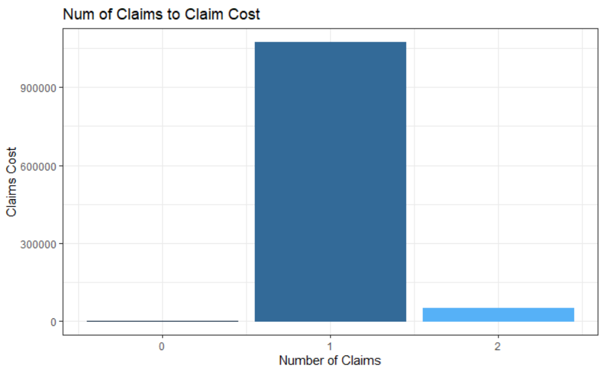
XGB Boost : XGB boosting models are distributed gradients - boosting decision trees that use decision tree ensemble learning methods . XGB boosting models combine both ensemble learning and gradient boosting methods to obtain the final model. Ensemble learning algorithms combine multiple machine learning models to obtain a better model. Gradient boosting uses multiple weaker models to generate a stronger model, before moving onto the next model gradient boosting provides a target outcome to reduce errors. Given the implementation of both ensemble learning and gradient boosting, XGboost is a highly accurate model that is commonly used by data scientists.

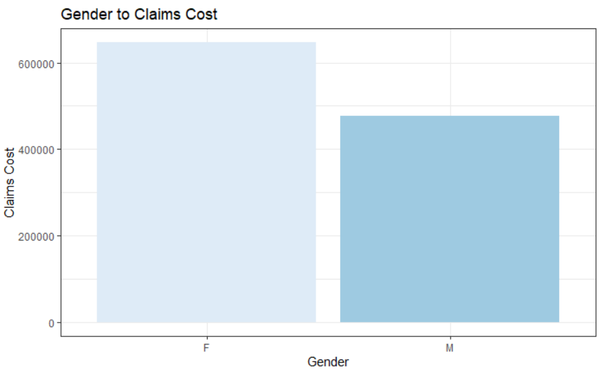
Tweedie: Tweedie distributions are a family of probability distributions which include the purely continuous normal and gamma distributions, the purely discrete scaled Poisson distribution, and the class of mixed compound Poisson – Gamma distributions which have positive mass at zero, but are otherwise continuous.

Poisson : Probability distribution that will be utilized in the models to predict response variables with counts of an event, for example in this case, the frequency a claim will occur. The poisson distribution will be utilized as the likelihood function for General linear models

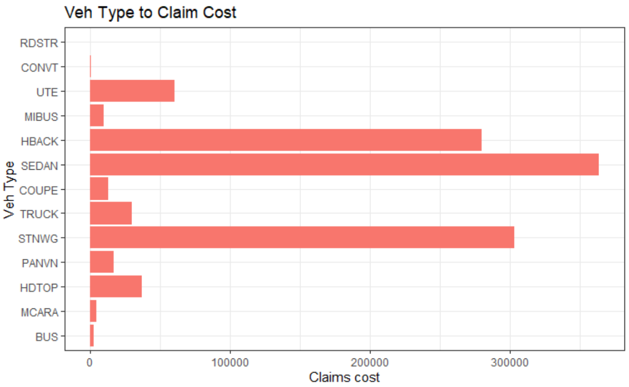
Gamma: Probability distribution that will be utilized in the selected models. The Gamma distribution is often used for modeling continuous outcomes that are non-negative. The Gamma distribution will be utilized as the likelihood function for general linear models. In this case we can use the gamma distribution to function in models predicting claim costs.

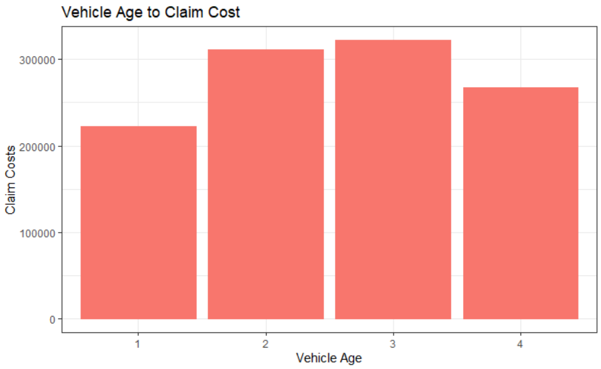
2.1 Exploratory Data Analysis



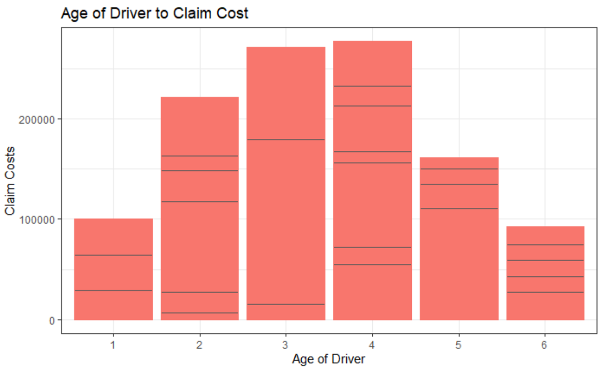


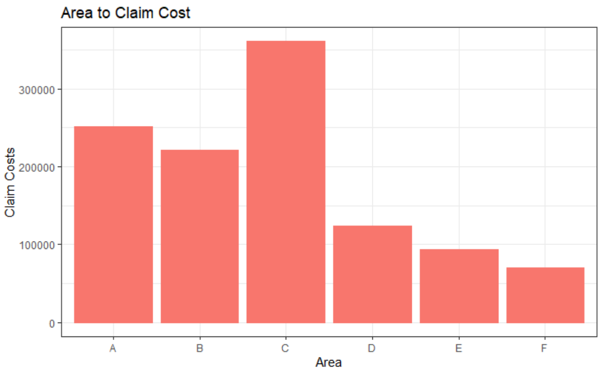
- Number of claims for a policyholder directly has an impact on total claim costs for a policy
- Female Drivers observed to have a increased claim costs compared to male policyholders





- Vehicles with type HBACK, SEDANS, and STNWG observe higher claim costs
- Vehicle age tends to have a normal distribution of claim costs varying between age 1-4





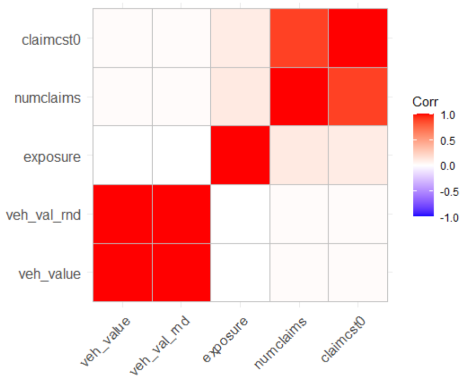
- Vehicle age tends to have a normal distribution of claim costs varying between age 1-6
- Areas for policyholder that live in C or A tend to have a higher total claims cost

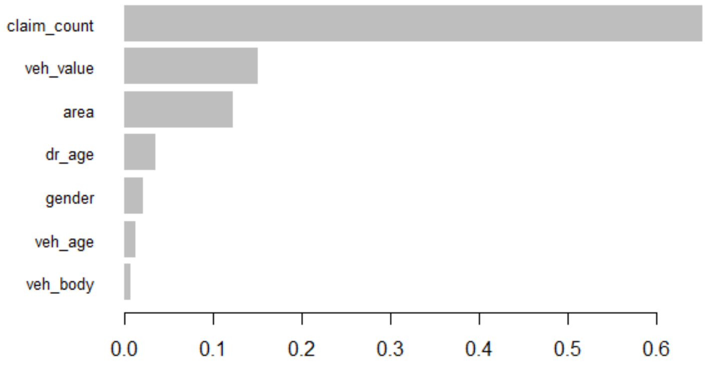
2.4 Model Performance

Model evaluation, to evaluate the models, the model with the highest gini coefficient on the hold set will be selected. The gini coefficient quantifies the amount of inequality that exists in a sample. The gini coefficient is a number between 0 and 1, where 0 represents perfect equality.

Utilizing model performance of root mean square deviation, coefficient of determinant of r^2, and mean absolute error

Feature selection when it came to models , we utilized a feature importance from the XGB Boosting model, to select features. Features like veh_body and veh_age observed the least importance.





2.5 Model Evaluation

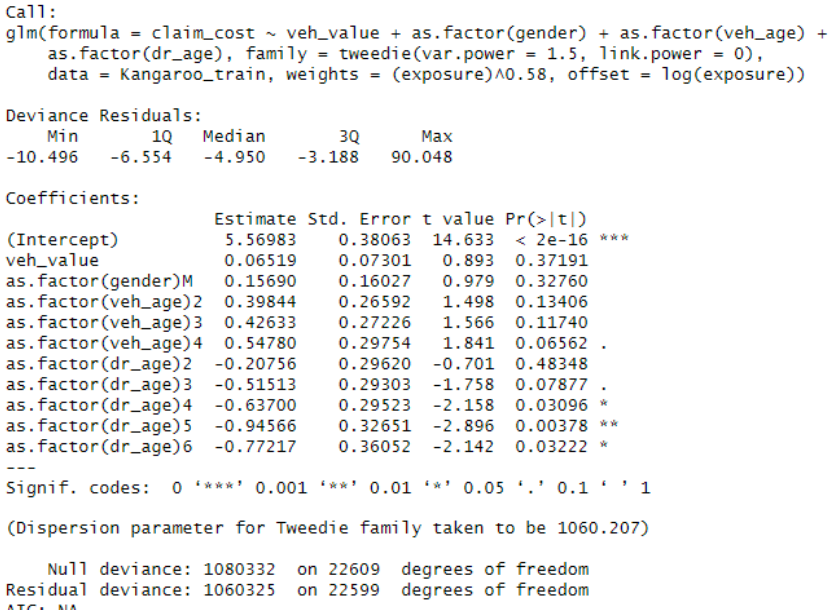
Various models were used in risk classification. Models were trained with 5 folds cross validation with 5 repeats. Comparing the models generated by performance indicators. Examining our model coefficients, I observed that the categorical features with the largest predictive power were the value of the vehicle and area of the driver when it came to predicting the target variables. Most of the models suffered from underfitting and were not able to capture relevant patterns, leading to less accurate predictions. Utilizing the validation set, we measured the model accuracy to find the best model.

Table 1. Experimental Results of Models for Regression and Classification

MODELS	**rmse**	**mae**	**r-squared**	**gini-coefficient**
Linear Model Simple linear model	1.600771e+02	2.864202e+02	3.366655e-04	0.1982134
Logistic Model Frequency model with Binomial distribution	1.231585e+03	1.600771e+02	6.893111e-04	0.09127537
Random Forest Claim Cost model with Binomial distribution	1.256579e+03	2.864285e+02	3.523062e-04	0.10127534
XGB Boost Pure Premium Model	1242.9111717	154.1452380	0.1849982	0.2903241
Tweedie GLM Claim cost model with Tweedie Distribution	1.266500e+03	1.701113e+02	1.895666e-03	0.3202927

GAM Claim Cost	1.256579e+03	2.864285e+02	3.523062e-04	0.1937518
GAM Claim Ind with Poisson Distribution	1.231682e+03	1.610054e+02	1.776744e-03	0.1039342
GLM Severity Model	6.240169e-04	2.722758e+02	6.240169e-04	0.1814456
GLM Frequency Model	0.6727654637	0.3159825249	0.0002522806	0.1863993

Tweedie GLM provided the best fit in classifying claim cost, loss cost, and severity. The final model can be



```
Call:
glm(formula = claim_cost ~ veh_value + as.factor(gender) + as.factor(veh_age) +
    as.factor(dr_age), family = tweedie(var.power = 1.5, link.power = 0),
    data = Kangaroo_train, weights = (exposure)^0.58, offset = log(exposure))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-10.496   -6.554   -4.950   -3.188   90.048

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.56983    0.38063  14.633  < 2e-16 ***
veh_value             0.06519    0.07301   0.893  0.37191
as.factor(gender)M    0.15690    0.16027   0.979  0.32760
as.factor(veh_age)2   0.39844    0.26592   1.498  0.13406
as.factor(veh_age)3   0.42633    0.27226   1.566  0.11740
as.factor(veh_age)4   0.54780    0.29754   1.841  0.06562 .
as.factor(dr_age)2   -0.20756    0.29620  -0.701  0.48348
as.factor(dr_age)3   -0.51513    0.29303  -1.758  0.07877 .
as.factor(dr_age)4   -0.63700    0.29523  -2.158  0.03096 *
as.factor(dr_age)5   -0.94566    0.32651  -2.896  0.00378 **
as.factor(dr_age)6   -0.77217    0.36052  -2.142  0.03222 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 1060.207)

    Null deviance: 1080332  on 22609  degrees of freedom
Residual deviance: 1060325  on 22599  degrees of freedom
```

As described above tweedie distribution provided the best performing model with its flexibility. While general additive models were flexible and provided high predictive power. XGB boosting models are very powerful even.

Conclusion

Several aspects of the results can be discussed, when it comes to Auto Insurance modeling variables closely related to the individuals information such as vehicle age, drivers age, and location provided to have a higher CC. If we were to collect information regarding the dataset, information like credit and tracking devices proves more valuable in predicting risk.

References

- Goldburd, M., Khare, A., Tevet, D., & Guller, D. (2016). Generalized linear models for insurance rating. Casualty Actuarial Society, CAS Monographs Series, 5.
- Effective Statistical Learning Methods for Actuaries I: GLMs and Extensions, Denuit, Hainaut, and Trufin, 2019
- Nathan Hubbel, Brendan Pham (HOST) . (2023,March 15). [Zoom Interview]. In discussion of predictive analytics for risk classification.