Master in Actuarial Science

Master's Final Work

Internship Report

Risk Modeling Journey - GLM and Impact Analysis

Jairo Moreira Caetano da Silva

Faculty Advisor: Alfredo Duarte Egidio dos Reis
Industry Supervisor: Drishti Singhvi
October-2021

Acknowledgments

I would like to thank the following people for their contributions:

Drishti Singhvi, industry supervisor, which helped me through all the stages of this work, defining topics of discussion, reviewing carefully the models and analysis, and explaining me both theoretical and practical subjects regarding the work with special attention and patience.
Alfredo Duarte Egidio dos Reis, faculty advisor, for all the comments, questions, suggestions and meticulous revision of this work, which were of central importance for the final result of this work.
Zhifeng Xu, friend and coworker, with clear explanations for several doubts I had during the work development, giving me support since the beginning and sharing excellent bibliographical references.
Derek Shupe, for all the assistance provided in order to produce this report.

Abstract

Generalized Linear Models (GLMs) are not a new topic. With increases in computing power and access to big data, actuaries have in fact been using GLMs in the insurance rating process for many years. Besides being well established, GLMs have straightforward interpretation which helps the communication with underwriters and the Product department.
Although many theoretical works have been done regarding GLMs in the insurance business, it is also important for actuaries to explore this subject in a work experience perspective. In this context, this project aims to present a practical view on the GLM model-building process from start to finish.
This work is applied to the Liberty Seguros motor insurance policy database and different risk models are built considering the automotive perils with higher exposure. The model evaluation step shows how well the model built aligns with historical data, which means that it is possible to verify the model ability to predict the risk behavior for new datasets after model development. During the impact analysis, it is shown the suggested changes in the prices for specific variables and how it affects the overall price, which can help the company to improve the underwriting process and profitability.

Keywords: motor insurance, GLM, impact analysis, risk modeling, insurance pricing

Resumo

Generalized Linear Models (GLMs) não são um tópico novo. Com o aumento da potência computacional e o acesso a *big data*, atuários têm de fato usado GLMs no processo de tarifação de seguros durante muitos anos. Além de ser bem consolidado, GLMs têm uma interpretação direta, o que ajuda a comunicação com subscritores e o departamento de Produto.
Apesar de muitos trabalhos teóricos terem sido realizados acerca de GLMs aplicados à indústria de seguros, também se faz importante para atuários explorar este assunto sob a ótica da experiência profissional. Neste contexto, este projeto tem como objetivo apresentar uma visão prática de todo o processo de construção de modelo de risco baseado em GLMs.
Este trabalho é aplicado à base de dados de apólices de seguro de automóvel da Liberty Seguros e são desenvolvidos diferentes modelos de risco considerando as coberturas automotivas de maior exposição. A etapa de avaliação do modelo mostra quão bem o modelo construído se alinha com os dados históricos, o que quer dizer que é possível verificar a capacidade do modelo em prever o comportamento do risco para novos conjuntos de dados após o desenvolvimento do modelo.
Durante a análise de impacto, são mostradas as mudanças sugeridas nos preços para variáveis específicas e como isso afeta o preço geral, o que pode ajudar a companhia a melhorar o processo de subscrição e a lucratividade.

Palavras-chave: Seguro de automóveis, GLM, análise de impacto, modelagem de riscos, tarifação de seguros

Contents

Chapter 1

Introduction

This is an internship report developed at Liberty Seguros Portugal, and consists of stages like fitting process, model validation, model evaluation, data manipulation and impact analysis.
The goal of this project is to explain how insurance risk models are built using Generalized Linear Models (GLMs), including the decision-making process and technical parameters evaluated when fitting and analyzing a risk model. The main objectives of this project are:

1. Show how the variables are selected to be part of a model, including decision making and rationalization.

2. Explain how the pure premium is calculated after building the risk model

3. Explain how the new model impacts the business and the necessity of adjustments to align with the company strategies

Chapter 2 covers the product book structure, data explanation and contextualizes the company and product which are objects of this project. Chapter 3 shows how the risk model is built, including fitting, validation and evaluation. Chapter 4 presents how risk scoring is done and perform the impact analysis.

The risk model of this study is a frequency-severity model, which separately models the claim frequency and average claim severity. The frequency is modeled assuming Poisson distribution and the severity is modeled assuming a Gamma distribution. The aggregate model is defined as a log-link

Poisson-Gamma model. The log-link function is adequate to a multiplicative model like the proposed one, because multiplicative models consider the interaction effects between the predictors, in contrast to additive models.

The risk model is composed of three perils modeled independently, RCM (Third Party Liability - Property Damage), RCC (Third Party Liability - Bodily Injury) and CCC (Crash, Collision and Rollover). Each peril is has two independent models: frequency and severity (i.e. RCM-F and RCM-S). The aggregate claims amount is obtained by multiplying the calculated factors for frequency and severity.

Chapter 2

Company, Product and Data

In this chapter we contextualize the company and product which are objects of this project, covering the product book structure and data explanation. The insurance company is Liberty Seguros Portugal, offering a Motor Insurance product in digital space, called *Liberty Sobre Rodas*. The product is commercialized by a company called Génesis, which belongs to the Liberty Mutual Group. We present the coverages considered in the modeling, the Bonus-Malus system and the variable categories.

2.1 Company and Product

Liberty Seguros has been in Portugal since 23 May 2003, through the acquisition of the former insurer Europeia from the Swiss group Credit Suisse. The company markets insurance solutions for private and non-life segments and currently has 533 employees.
Liberty Seguros is the 6th player in the Portuguese motor insurance market. It has more than 130 million euros in written premium, corresponding to 7.1% of the national market share in 2020, according to Autoridade de Supervisão de Seguros e Fundos de Pensões (ASF).
Liberty Sobre Rodas was already implemented in Spain and will pass through some adaptations for the Portuguese market. However, all the risk modelling is being done specifically according to Liberty's experience of historical claims in the Portuguese Market. Nowadays, Liberty Seguros operates in Portugal only through the broker channel, which means they do not have a digital

channel to connect directly with the final client. The product is designed for individual, private and light vehicles, and includes basic and comprehensive coverages.
The main benefit of this product is on the practicality for policy quotation. With only five direct questions the client can obtain a quote for his motor insurance policy through an online channel, without the need of interact with a broker or any other attendant. The motor insurance is a great product to start this transition to the digital quotation due to its demand in the Portuguese market. Since it is an obligatory insurance, the big volume of cars in the country makes this product an excellent opportunity for client base expansion and improve company's revenue.

2.2 Coverages

In Table 2.1 we can see the coverages considered in the model, showing their original abbreviated name in Portuguese and the corresponding translation to English.

Portuguese Abbreviation	Coverage Name In English
RC	Third Party Liability (TPL)
RCM	TPL - Property Damage
RCC	TPL - Bodily Injury
CCC	Crash, Collision and Rollover

Table 2.1: Coverages abbreviation with description

The choice of coverages was done according to the relevance of the coverage in terms of aggregated loss. The other main covers, like Windscreen and Theft represent a small amount of loss when compared with the others chosen.
Third Party Liability – Property Damage: Losses ensuing from damage or injury to movable or immovable property or animals who, as a result of an incident covered by this contract, suffers damage or injury that entitles them to compensation or indemnification under the terms of civil law and this policy.
Third Party Liability – Bodily Injury: Losses ensuing from injury to physical or mental health to anyone who, as a result of an incident covered by this contract, suffers damage or injury that entitles them to compensation or indemnification under the terms of civil law and this policy.

Crash, Collision and Rollover: The following definitions apply for the purposes of this cover: Crash - when the vehicle strikes any other stationary object or the vehicle is struck while stationary; Collision - when the vehicle strikes any other object in motion; Rollover - when the vehicle is no longer in its normal position, but not as the result of a crash or collision.

2.3 Bonus-Malus System

For Liberty Sobre Rodas, we consider the No Claim Discount (NCD) to apply the Bonus-Malus system. In Table 2.2, we can see the Bonus-Malus levels and premium percentages for RC and CCC.

Level	Years Without Claim	TPL	CCC
1	9	45%	45%
2	8	45%	45%
3	7	50%	50%
4	6	55%	55%
5	5	60%	60%
6	4	65%	65%
7	3	70%	70%
8	2	80%	80%
9	1	90%	90%
10	0	100%	100%
11		110%	110%
12		120%	120%
13		130%	130%
14		150%	150%
15		180%	150%
16		250%	150%
17		325%	150%
18		400%	150%

Table 2.2: Bonus-Malus levels for TPL and CCC

	Next Level If # Claims in the year						
Current Level	0	1	2	3	4	5	6+
1	1	4	7	10	13	16	18
2	1	5	8	11	14	17	18
3	2	6	9	12	15	18	18
4	3	7	10	13	16	18	18
5	4	8	11	14	17	18	18
6	5	9	12	15	18	18	18
7	6	10	13	16	18	18	18
8	7	11	14	17	18	18	18
9	8	12	15	18	18	18	18
10	9	13	16	18	18	18	18
11	10	14	17	18	18	18	18
12	11	15	18	18	18	18	18
13	12	16	18	18	18	18	18
14	13	17	18	18	18	18	18
15	14	18	18	18	18	18	18
16	15	18	18	18	18	18	18
17	16	18	18	18	18	18	18
18	17	18	18	18	18	18	18

Table 2.3: Transition rules for TPL and CCC Bonus-Malus Levels

The transition rules are defined as follows:

1. If the driver doesn't have a claim in the last year, he goes down 1 level.
2. If the driver has one or more claims in the last year, he goes up 3 levels for each claim.
3. The starting level is 10, the minimum level is 1 and the maximum level is 18.

In Table 2.3, we can see the transition rules matrix for B-M levels applied for both RC and CCC.

2.4 Variables Categories

The variables defined from our dataset are clustered in different types. In the model built, we considered: time information (TI), driver information (DR), policy information (SA), vehicle information (VH), geographical information (GEO) and claims history (CL). Regarding the type of fit, we considered: simple factor (SF), custom factor (CF) and variate (VR). The list of variables fitted in the risk model, their categories and fit types are shown in Appendix B.1.

Chapter 3

Risk Modeling Process

In this chapter we can see how the risk model is built, going through the fitting process, explaining the statistical tools used in Emblem® and the thought process, showing examples of variables fitted, validating the model using a holdout dataset and evaluating the model with two different methodologies.

3.1 Pure Technical Models

To build the risk models, we need to train and validate the generalized linear models. For the Pure Technical models, we divide our dataset in two different sets:

1. Training dataset: 70% of the data, used in the model training, to estimate the best coefficients for the model regressions.
2. Hold out dataset: 30% of the data, used in the model validation, to check if the regressions obtained in the training phase are consistent with a different dataset.

The choice of what data (policies) will be used in each of these parts is made completely random. For this reason, we need to create a variable called "Random20" and this variable tells us, for each row of the database, if this policy will be used in training or hold out. The *Random20* variable works in the following way: we assign a random number (from 1 to 20) to each of the policies in the database and, if the assigned number goes from

1 to 14, this policy will be considered in the training data used to build the regressions. However, if the number assigned goes from 15 to 20, this policy will be considered in the holdout data used to validate the regressions.

3.1.1 Fitting Process

After creating a separate file containing just the training dataset and defining which variables are going to be fitted in the model, we start the recursive fitting process. The recursive fitting process is described in the following steps:

1. Fit the variable and analyze its contribution to the model
2. Check the statistics, graph, standard errors, and standard error matrix to see if the variable and all its levels are significant
3. Create an interaction between the fitted variable and *Random04*[1], to analyze the trends of the variable in four random subsets, for consistency purposes
4. Check the effect of the fitted variable in the other variables fitted previously. If it makes another variable insignificant, we remove the new variable from the model.
5. If all the previous checks are good, we consider that the added variable improves the quality of the model and we keep it in the model. If it fails in the previous checks, we remove it.
6. Set a new Reference Model if the new variable was included in the model
7. Go to the next variable

3.1.2 Statistics and Graphs

In order to explain how we interpret the statistics and graphs, Examples 1 and 2 are presented.

[1] A variable that works similarly to *Random20*, but dividing the data in four different subsets

Example 1: Variable V_{17}, RCM-F

Table 3.1 is used to analyze the addition of a variable in the model. In this case, we are using the model for the peril RCM-F (frequency) after fitting the variable V_{17}. The table presents three columns:

1. Current model: the model after the inclusion of the variable that is being fitted in the moment
2. Reference model: the model that considers all the variables already fitted prior to the current variable being fitted
3. Difference: the difference between the two models for the statistics considered in the analysis for keeping or not the variable in the model

Model Label	Current Model	Reference Model	Difference
Fitted Parameters	63	61	2
Deviance	965,109.5	966,986.4	-1,876.9
Chi-Square Percentage		Sub-Model	0.0%
AICc	1,003,307	1,004,155	-848

Table 3.1: Statistics Comparison

Now, it is important to detail the interpretation of the statistics values in the column "Difference":

1. Fitted parameters: Since V_{17} is fitted as variate due to its nature, the value "2" means that this variable was fitted as a polynomial of second order, reducing two degrees of freedom in the model.
2. Deviance: The deviance reduction after including this variable is massive. It means that this variable is a good predictor for the current model.
3. Chi-Square Percentage: The value of 0.0% means that the variable reduction is in fact significant for any significance level.
4. AICc: The AICc reduction is also big, which means that the addition of this variable improves the quality of the model, considering both goodness of fit and the simplicity of the model.

Besides checking the statistics, another way to evaluate the goodness of fit of a variable is by analyzing its graph. Figure 3.1 shows the observed average, the fitted average considering all the variables in the model so far, and the model prediction at base levels, which represents the behavior of the variable across its levels for the reference risks.

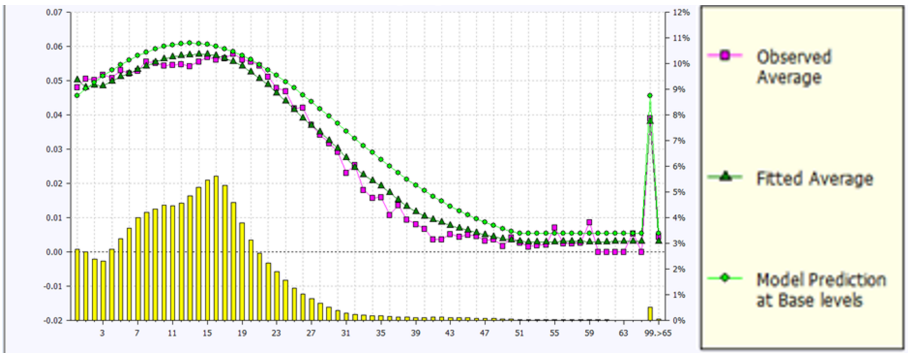

Figure 3.1: V_{17} fitting graph

For variable V_{17}, we can see that:

1. Observed Average and Fitted Average are really close to each other, which is a good sign.

2. The model prediction at base levels have basically the same trend of observed data, which means that if two different policies have the same characteristics for all other variables, any difference in V_{17} will lead to a good estimate of the difference in the risks of these two policies.

Another useful graph analysis is the "random 4" comparison. In this analysis we check the interaction between V_{17} and the variable $Random04$, used to divide the full dataset in four random subsets and see if the trend is the same for the four different groups, which is a strong sign that the variable follows this specific behavior. Figure 3.2 shows the interaction between V_{17} and $Random04$.

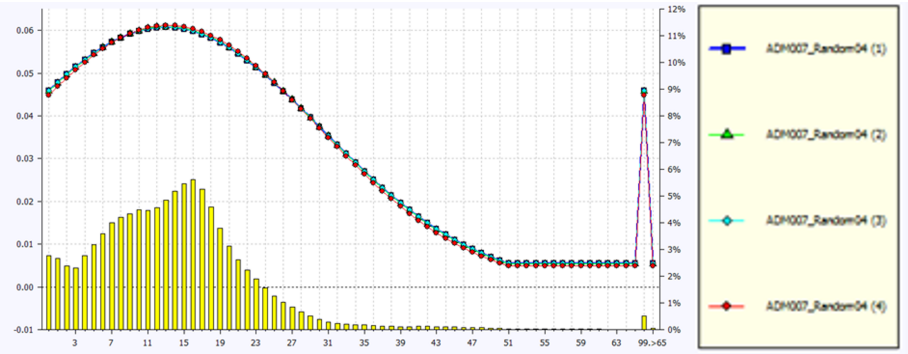

Figure 3.2: V_{17} fitting graph

By analyzing the graph, we see that all the four subsets follow precisely the same trend, which is a good indicator of the variable expected trend for future predictions.

Example 2: Variable V_{03}, RCC-S

Variable V_{03} is an example of a variable that we didn't keep in the RCC-S (severity) model. As V_{17}, it is a numeric variable and we tried to fit it as a variate. The following tables show the results of this iterating process.

Statistics after fitting the first order:

Model Label	Current Model	Reference Model	Difference
Fitted Parameters	50	49	1
Deviance	30,289.56	30,294.32	-4.76
Chi-Square Percentage		Sub-Model	32.6%
AICc	112,785.2	112,784.1	1.1

Table 3.2: V_{03} statistics comparison, first order

After fitting the first order, we can see that the deviance reduction is small, suggesting that the variable is not of great help on explaining the observed data response. Chi-Square Percentage shows that the variable is not significant at 5% level. So, we conclude that the first order polynomial is not a good fit for this variable, and we try to fit the next order.

Statistics after fitting the second order:

Model Label	Current Model	Reference Model	Difference
Fitted Parameters	51	49	2
Deviance	30,265.55	30,294.32	-28.77
Chi-Square Percentage		Sub-Model	5.4%
AICc	112,780.5	112,784.1	-3.6

Table 3.3: V_{03} statistics comparison, second order

We can see that the second order had a better result according to Chi-Square Percentage and AICc but was not significant at 5% level. The Chi-Square Percentage was 5.4% so we reject the hypothesis that the variable reduces deviance in the model, and we decide to fit an additional order.

Statistics after fitting the third order:

Model Label	Current Model	Reference Model	Difference
Fitted Parameters	52	49	3
Deviance	30,262.37	30,294.32	-31.95
Chi-Square Percentage		Sub-Model	9.1%
AICc	112,782.0	112,784.1	-2.1

Table 3.4: V_{03} statistics comparison, second order

As seen after fitting the third order, the addition of a new order always reduce deviance. However, in this case it doesn't represent an improvement in the model according to the AICc criteria and Chi-Square percentage increase, which makes us keep the variable out of the model.

To corroborate our decision, we can analyze the second order fitting graph in Figure 3.3.

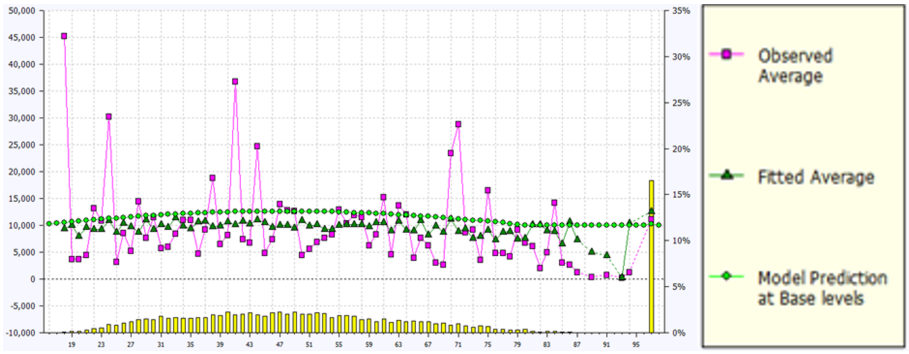

Figure 3.3: V_{03} fitting graph, second order

In the graph for the second order, we see that observed data doesn't show a clear trend itself and the fitted data can't capture the variable's behavior. In this case, the better approach is to keep the variable out of the model.

3.1.3 Standard Error and Standard Error Matrix

The level of the statistical difference between the two parameters is indicated by the color of the font. Numbers less than 50% are highlighted in green, those from 50% to 75% are highlighted in grey, and those above 75% are highlighted in red. This works for both Standard Error Percentage and the Standard Error Matrix.
The standard error matrix percentages represents a measure of the covariance of two different levels of a variable. The "standard error of the parameter difference" percentages can be used to aid simplification of the model. A high percentage indicates little statistical difference between the two parameters and, therefore, that the rating factor levels may be grouped in subsequent fits.
In order to explain how we interpret the standard error and the standard error matrix, Examples 3 is presented.

Example 3: V_{05}, RCC-S

Table 3.5 shows basic statistics of V_{05}. We can analyze the column Standard Error to check if the variable is significant.

	Name	Value	Standard Error (%)	Weight (%)	Exp(Value)
-	L1			71.4	
19	L2	0.0353	233.1	15.1	1.036
20	L3	0.2230	78.8	2.7	1.2498
21	L4	0.2580	55.0	10.7	1.2944
22	L5	-1.8874	68.3	0	0.1515

Table 3.5: V_{05} basic statistics

When we try to fit variable V_{05} in the RCC-S peril, all the levels are not significant, according to the standard error check. All standard error percentages are over 50% so we consider that the levels are not significant different from the reference level.

This assumption is ratified by the standard error matrix. In Table 3.6 is shown the covariance between the variable levels.

	L1	L2	L3	L4	L5
L1					
L2	233.1				
L3	78.8	99.9			
L4	55.0	65.5	626.2		
L5	68.3	67.1	61.6	60.3	

Table 3.6: V_{05} standard error matrix

We can see that any level is significantly different than any other in the variable V_{05}, which confirms that the variable is not significant for the model.

3.1.4 Fitting Variables

In this section is shown Examples 4 and 5, which demonstrate how we fit different types of variables like numeric and categorical variables.

Example 4: Variable V_{17}, RCM-F

The variable V_{17} is a numeric variable. Our first approach is to fit a variate as a polynomial of the first order and analyze the fit through the graph. Since our data don't have much exposure after level 50, we decide to truncate the

variable after this level.

Figure 3.4 shows the first degree fit for V_{17}:

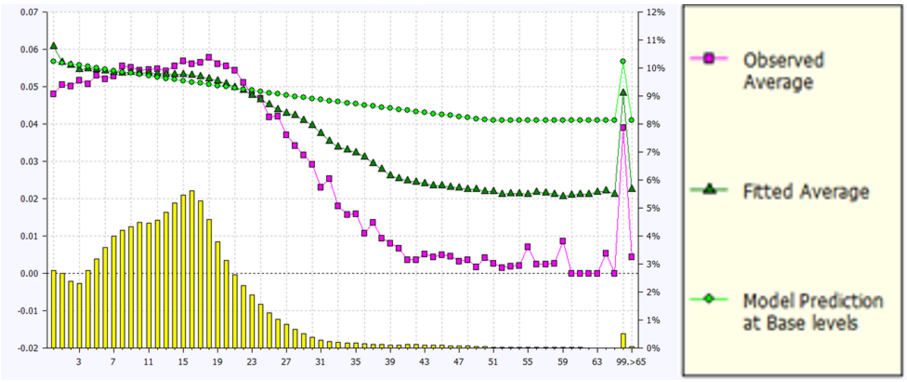

Figure 3.4: V_{17} fitting graph, first order

Fitted Average is very far from Observed Average, which suggests that the fit is not good enough. So, in this case, we can try to add the second order and check the fit again.

Figure 3.5 shows the second degree fit for V_{17}:

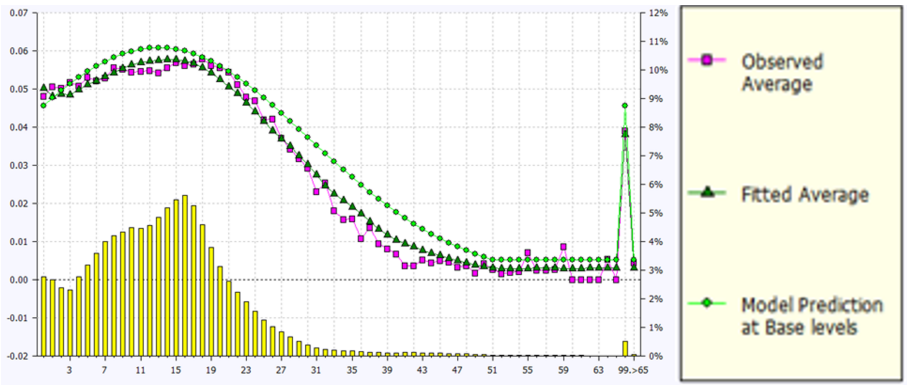

Figure 3.5: V_{17} fitting graph, second order

Fitted Average and Observed Average are close and following a similar trend. Model Prediction at Base levels follow a similar trend to Fitted and Observed data, which is desired. This visual analysis suggests that the addition of the second degree is relevant for the model prediction.

In order to confirm if the second order fit is good enough, we fit the third order and check if there is significant improvement. Figure 3.6 shows the third degree fit for V_{17}:

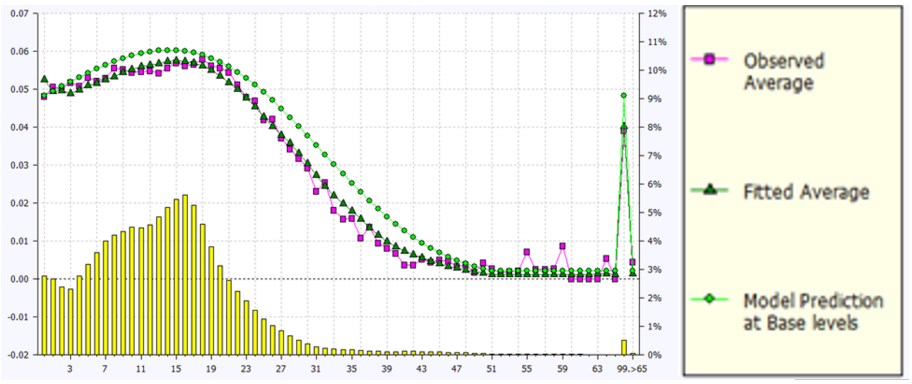

Figure 3.6: V_{17} fitting graph, third order

Since our previous fit (second order) already gets a great fit according to our observed data, the 3rd degree doesn't show any remarkable improvement. It is expected a deviance reduction after the addition of this term, but it doesn't result in a much better prediction. There is no reason for us to keep the 3rd degree term, avoiding overfitting the model and following the principle of parsimony.

Example 5: Variable V_{20}, RCM-F

Variable V_{20} is a categorical variable. Our first approach is to fit a simple factor and check for not significant levels in the standard error matrix. It is a relevant sign that we should group levels together to make all level groups significantly different in the end.

	L1	L2	L3	L4	L5	L6	L7	L8
L1								
L2	49.3							
L3	34.2	657.6						
L4	9.7	14.5	9.7					
L5	38.6	148.0	142.1	13.2				
L6	63.1	11,062.8	838.0	19.1	175.8			
L7	3.9	4.9	2.9	5.9	6.0	6.8		
L8	130.6	366.3	405.1	78.8	1,175.8	369.1	26.3	

Table 3.7: V_{20} standard error matrix, before grouping levels

After fitting a simple factor for the variable V_{20}, we can see from Table 3.7 that several levels are not significantly different from some others (not green). In this situation, we can assume that their rating factors are close, so we group these levels together and get an average rating factor for the group. In this case, level 8 stands for "Unknown" value, so we group it with the highest exposure level, which is level L4.

The result for this grouping step can be seen in Table 3.8.

	L1	L2	L3	L4
L1				
L2	33.7			
L3	9.7	9.4		
L4	3.9	2.8	5.9	

Table 3.8: V_{05} standard error matrix

After grouping the levels with similar rating factors, we have four different groups, all significantly different from each other, which represents a good fit.

We can also confirm our conclusion by analyzing the fitting graph for variable V_{20} in Figure 3.7.

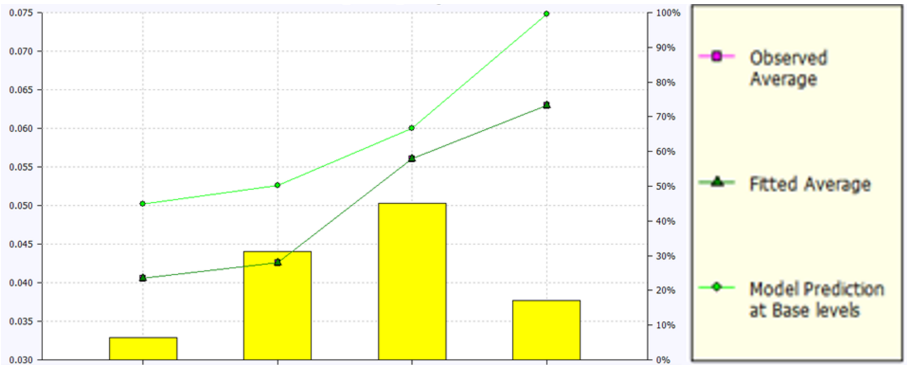

Figure 3.7: V_{17} fitting graph, third order

After the grouping step, we can see that the fitting for this variable is pretty accurate. The Fitted Average and Observed Average are overlaid, and the model prediction trend is going in the same direction as the Observed Average data.

3.2 Holdout Validation

After fitting the training model, it is important to validate this model in a new data sample. This validation is done using the holdout data. In this phase, we do the following steps using the 30% data Emblem data file:

1. Adapt the model: this functionality imports all the fitting done for all the variables from a desired model. By doing this, we will replicate the training model in the 30% data from holdout. In this case we adapt from the training model, automatically fitting the same variables as in this model to the new dataset.

2. Check standard errors: we need to check if all the variables remain significant in the new dataset. To check this, we can see the new standard errors for the holdout dataset and check if all the levels are significant for simple and custom factors and each polynomial order for variates.

In the Table A.2 in the Appendix, we can see, for RCM peril, the table "Fitted Parameters" from Emblem. It shows the most relevant information about the fitted variables, including their standard error percentages in the

holdout dataset and their relativities in the last column.
From Table A.2 we can see that all levels are significant in the model so we can conclude that the model is consistent with the new dataset.

3.3 Restricted Model

Restricted models are models where one or more restrictions are imposed according to the nature of product offered. In this case, the product offered is restricted only for:

1. Individual clients, excluding companies that would like to contract motor insurance for a car fleet.

2. Light vehicles, excluding trucks and trailers and heavier vehicles.

3. Vehicles used for private purposes, not commercial activities.

4. New business, not available for policy renewals.

Therefore, it is necessary to take this in consideration when the models are built, once the database include risks with all types of characteristics and not only the risks of customers to whom the product will be offered.
After fitting the training model without restrictions and validate it on the hold-out data, it is time to build the restricted model, imposing the necessary restrictions for the product. The steps to get the restricted model, which will be used in our scoring process, is the following one:

1. **Zeroweight Data**: we start from the 70% training Emblem data and apply the restrictions mentioned before by zeroweighting the levels not relevant for our product. Zeroweight is the process of applying weight zero to specific levels in a variable. After doing that, our data is filtered according to the product needs and our model will be fitted considering only this part of the data.

2. **Adapt Model**: since we already have the filtered data, now we adapt the model. This step consists in load the unrestricted fitted model in the current data file (restricted data). The result of this step is to have all the same variables fitted in the restricted data, the same way they were fitted before.

3. **Offset Model**: after adapting the model, the next step is to offset the model. The aim of offsetting the model is to fix the variables fitting into the new model.

4. **Offset Bonus-Malus to booklet**: when the bonus-malus variable was fitted in the unrestricted model, it was fitted as a simple factor, assigning a different relativity to each original level. Now we substitute the bonus-malus free fit by the discount/penalty table for the product, fixing the values the same way they impact the final pricing in the bonus-malus system.

3.4 Model Evaluation

Model evaluation is a very important component in the modeling process. In this stage, we can use different methodologies to evaluate how good the new predictive model is, comparing it with the historical data and previous models. In order to assess the proposed model, we use two different graphical tools: the Lift Chart and the Gain Curve.
The Single Lift Chart is used to assess the performance a model. We use the average response of the raw data as our reference model. The other is the new model, which we want to evaluate. The Gain Curve is an evaluation curve that assesses the performance of the models and compares the results with the random selection.
We choose to do the following analysis using the RCM peril because it has more data than other perils, which helps the accuracy of the models built.

3.4.1 Single Lift Chart

The Single Lift Chart is created using the following steps:

1. Sort the data from lowest to highest based on its response values.

2. Group your data in bins based on Step 1 and putting similar volume of exposure in each bin. We use twenty bins, but other values can be used.

3. For each bin, calculate the average response value (blue triangles) and the average actual value (red squares).

4. Plot the results. On the horizontal axis we have the bins (from lowest to highest) and on the vertical axis we have the average response value and the average actual value.

In the following two graphs, we will analyze the Single Lift Chart for the RCM peril, for Frequency and Severity models, respectively.

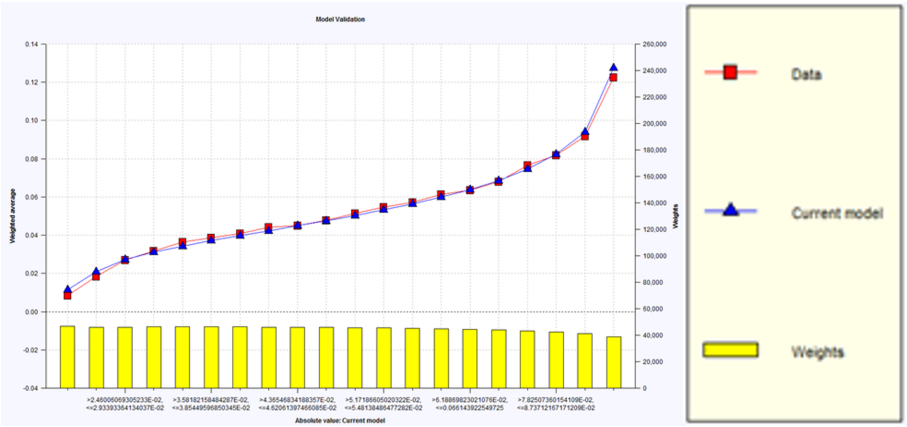

Figure 3.8: RCM-F single lift chart

Since the current model is close to the data in the hold out sample, we can conclude that the model has a good response.

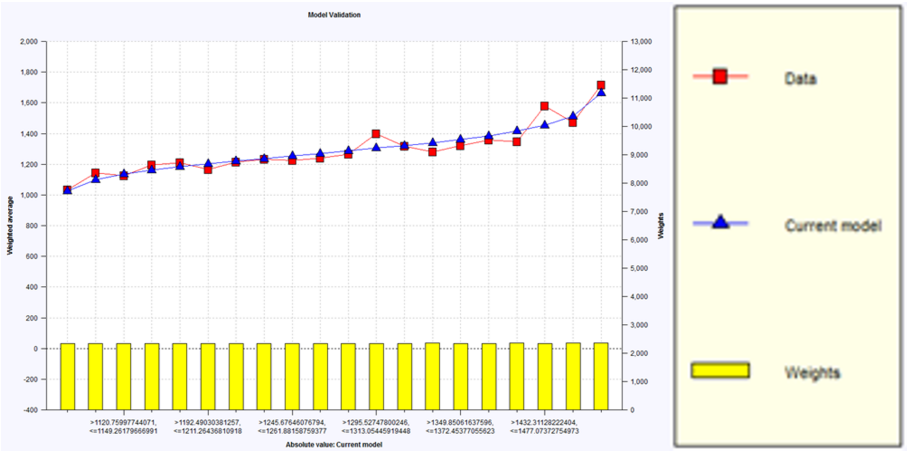

Figure 3.9: RCM-S single lift chart

For the severity model, we have the same analysis as the frequency model, but the final result is not as close to data as in the frequency model since the severity model has much less data, which usually reduces the model accuracy.

3.4.2 Gain Curve

Model 1 (green line) is the new holdout model, Model 2 (red line) is the old holdout model. The Reference (blue line) is a straight line that represents the random selection. We use holdout models to generate the Gain Curve because it will show how well the model is generalizing. Using the training model over the training set would present an over optimistic result, which is not desired.
In the following two graphs, we will analyze the Gain Curve for the RCM peril, for Frequency and Severity models, respectively.

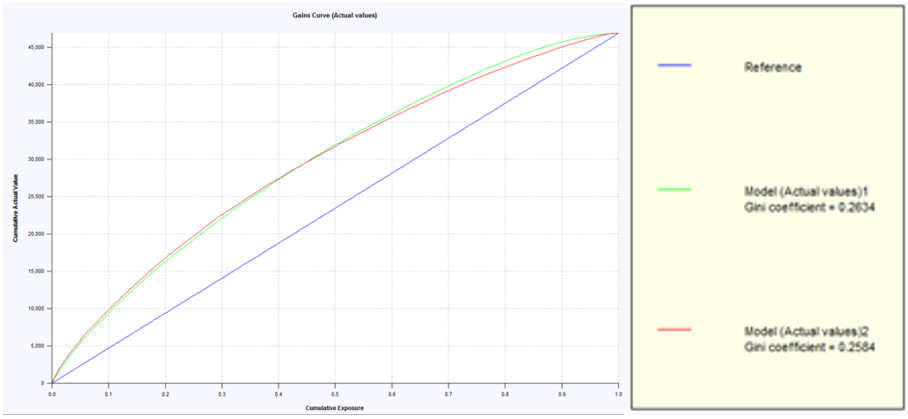

Figure 3.10: RCM-S single lift chart

From the Gain Curve we can conclude that the models are similar, and the Gini coefficient is a bit higher for the new model (0.263 and 0.258), which represents a slightly improvement when comparing the new model against the old one.

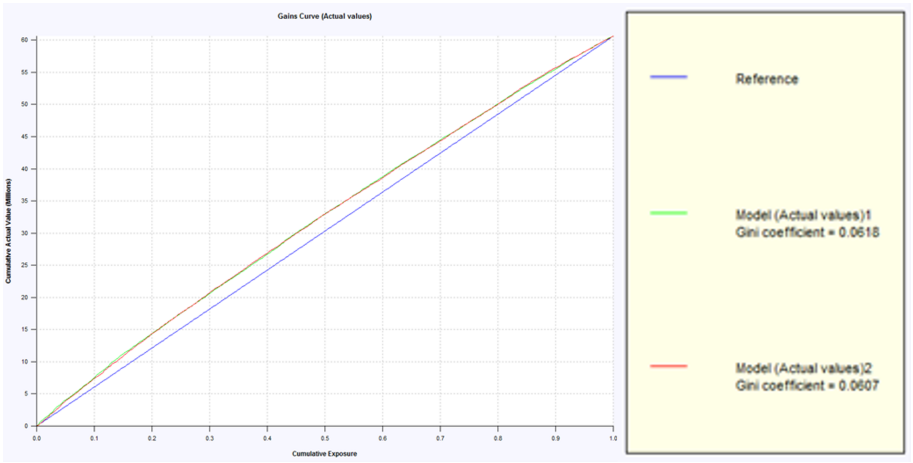

Figure 3.11: RCM-S single lift chart

The models are really close in this case so we could say that both models have similar response. The Gini coefficients are considerably lower in the severity models (0.062 and 0.061) when compared to frequency models (0.263 and 0.258). This can also be explained by the data volume available for frequency being much more robust than the data available for severity models.

Chapter 4

Risk Scoring and Impact Analysis

After building and validating the model, the following steps are due to produce the risk scoring and perform an impact analysis.
Risk scoring is the process of calculating a number that reflects the severity of a risk due to specific factors. In this case, the factors come from the variables selected in each of the six risk models built (two for each peril: frequency and severity). Since this is a multiplicative model, each level in each variable has an exponential factor, called relativity. For each record in the historical data, according to the value of each variable for that risk, a relativity will be assigned to that risk and the calculated scoring risk will be given by the multiplication of all assigned relativities in that peril, considering both frequency and severity. The result of this multiplication is called the Burning Cost Premium for that risk, which is the annual claim expected value.
The Impact Analysis is used to identify the potential consequences of a change or estimating what needs to be modified to accomplish a change. After the process of risk scoring, it is necessary to verify the impacts of the new model in contrast with the old model, considering different aspects like price dislocation, relativities, and conversion rates, which can drive changes in the proposed model.

4.1 Risk Scoring

In order to perform the risk scoring, we use the Radar® software, developed by Willis Towers Watson. The Radar project is used to process all the historical information from the policies, scoring based on the risk models built and generate the outputs for the impact analysis. The image below shows the process flow in this project.

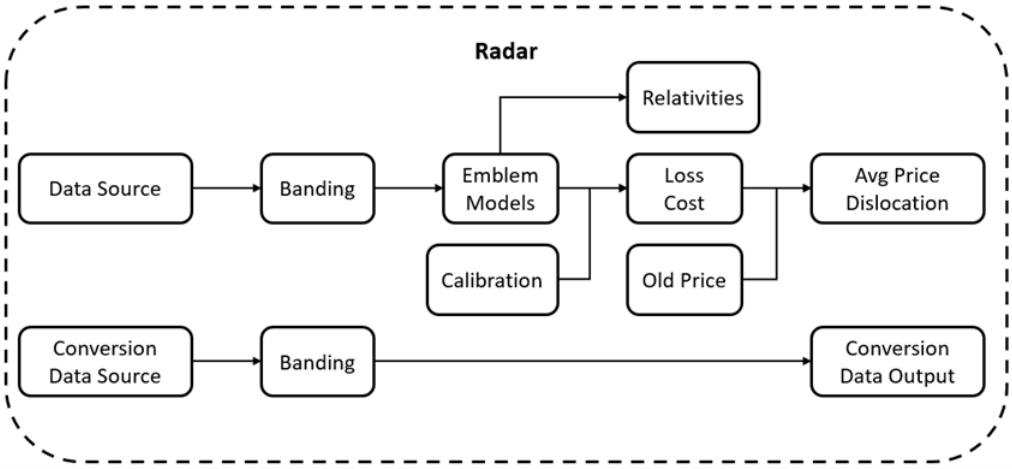

Figure 4.1: Radar process flow

The Data Source is a database from SAS, and for all the policies in the restricted scope, it has the values for each one of the variables used in the model. This information will be used to assign the corresponding relativities and calculate the burning cost.
Banding is the process of grouping levels for each variable the same way as it was done in the Emblem models. In this step, we define the upper and lower bounds for each level in numeric variables and assign the same level for different values in categorical variables. The Emblem Models present the relativities for all the levels of each variable. In this project we have six Emblem Models, which are frequency and severity models for three different perils: TPL-Property Damage, TPL-Bodily Injury and Own Damage. In this step, all the banding data will be an input for the Emblem Models and will generate an output called "Predictor", which is the multiplication of the base factor with the relativities for a given policy, since it is a multiplicative model. Each of the six Emblem Models will have its own Predictor that will

be used in the calculation of pure premiums. From these Emblem Models, we can export all the model relativities used in the scoring process. These values will be used in the impact analysis graphs to analyze the impact of each variable exclusively. The relativities are treated separately and go through the following steps:

1. Calculation of the overall relativity: in order to get an overall relativity combining all the perils, we calculate it by multiplying all the factors and applying weights to each peril according to their weight in the old model.

2. Recalculate overall relativities after rebanding variables: since the banding in the models are different from the final banding used in the impact analysis for several variables, it is necessary to recalculate the relativities for the new levels. In this case we apply weighted average to consider both the relativity and the exposure of the old bands that will compose the new band. This situation often happens for numeric variables with too many levels, which is not good to see in the graph at the same time.

3. Rebase relativities for the reference level: for analysis purposes, we rebase the relativities and set as reference the level with higher exposure (relativity = 1). To do that, we divide all the relativities by the relativity of the reference level, which gives relativity 1 to the reference level and keep the proportion between all the levels.

In order to match the loss ratio of the new model with the old model and be able to clearly compare the differences between the models, we calculate calibration factors. These calibration factors are calculated for each peril, which means that the total premium for each peril will match after calibration. The factors are calculated as the sum of all premiums in the previous model divided by the sum of all premiums in the new model, as shown in the following formula:

$$Calibration.Factor = \frac{\sum Premium.Old}{\sum Premium.New}$$

After calculating the calibration factor, it is included as multiplying factor in the premium calculation.
In the Loss Cost stage, we calculate the burning cost premium per peril by

multiplying all the relevant factors. For instance, the TPL-PD burning cost premium is calculated as shown in the following equation:

$$Burning.Cost_{TPL.PD} = ID_{TPL.PD} \times FREQ_{TPL.PD} \times SEV_{TPL.PD}$$

$$\times Calibration.Factor_{TPL.PD}$$

Terms of the equation:

1. "ID" is an identifier that says if the given policy has coverage for the corresponding peril
2. "FREQ" is the frequency component of the premium for the corresponding peril
3. "SEV" is the severity component of the premium for the corresponding peril
4. "Calibration.Factor" is the calibration factor used to match the sum of premiums between the models for the corresponding peril

The old price is calculated for all the policies in the dataset using past relativities, in a similar way as in the new model. The premiums are calculated for the same perils and will be used for calculating the calibration factors and for comparison in the impact analysis.
The output file presents for all the policies: the banded level of each variable, the calculated premiums for the new model and the premiums for the old model. The values of each variable are used to calculate the exposure of each level in each variable while the old and new premiums are used for calculating the average price dislocation.

4.2 Impact Analysis

The impact analysis is a process to evaluate the impact of the new scoring models when compared to the old model. It is relevant for the product team to evaluate the proposed impacts on price dislocation, relativities, and conversion rates. The analysis is composed by the following three graphs:

1. **New Model vs Old Model Relativities**: this graph shows the relativities for the new and old model for each level of the variable and

its exposure in millions. It is a good way to understand the impact of a single variable in the price for each exposure band. If the new model has higher relativity than the old model for a specific band, it means that the model is forcing the price up for this variable and level. It doesn't consider the impact of the other variables in this same layer of exposure.

2. **Average Price Dislocation**: this graph shows the average price dislocation in percentage for each level of the variable and its exposure in millions. The price dislocation is calculated as:

$$Price.Dislocation = \frac{\sum Premium.New}{\sum Premium.Old} - 1$$

It is a good way to understand the impact of all the fitted variables in the price for each exposure band. If the price dislocation is negative for a given band, it means that the new model is providing a price lower than the old model for this specific level. In this case, it considers the impact of all the variables fitted in the model for this layer of exposure.

3. **Historical Conversion Rate**: this graph shows the conversion rate for each level of exposure for a given variable and the proportion of quotes in that same level. The conversion rate is calculated as the sum of all converted quotes divided by the sum of all quotes, as shown in the following formula:

$$Conversion.Rate = \frac{\sum Conversions}{\sum Quotes}$$

An insurance quote is an estimate of how much a new insurance policy will cost (price quote) while the conversion is the act of converting a quote into a current policy (turn the potential client into a real client). It is a good way to understand the interest of the potential clients in the product. Since clients are strongly driven by price in motor insurance, there is a high and positive correlation between price and conversion rate. So, if the conversion rate is low for a specific level, a possible strategy is to decrease the price for that level to try to capture a bigger share of this market.

To illustrate the impact analysis, variables V_{03} and V_{18} were selected. For both variables, we show the graphs and the interpretation for each graph in

the impact analysis.
Impact Analysis for V_{03}:

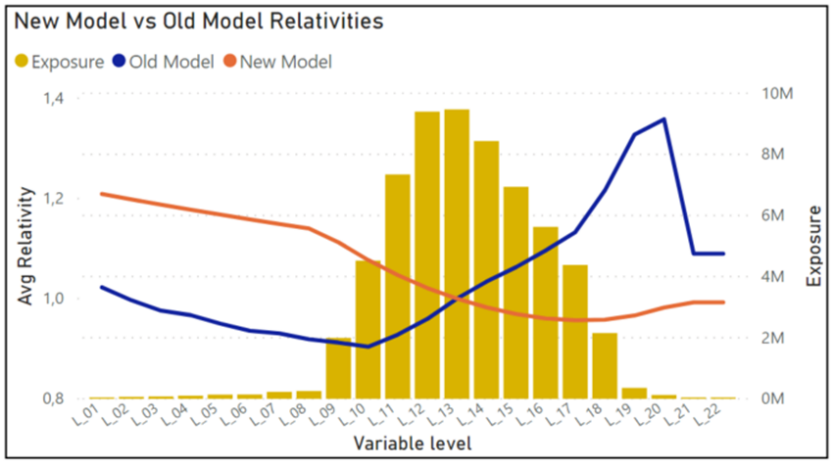

Figure 4.2: V_{03} Relativities graph

Relativities: between levels L_{01} and L_{06}, the new model suggests an increase of the relativities compared to the old relativities, which impact prices positively.

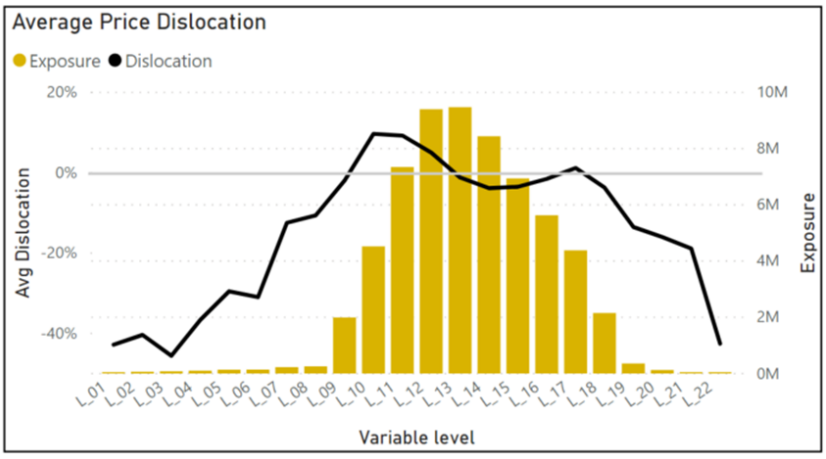

Figure 4.3: V_{03} Average Price Dislocation graph

Price Dislocation: even increasing the relativities in these bands, the overall price is decreasing for levels between L_{01} and L_{06}, which can be explained as an effect of other variables in the model.

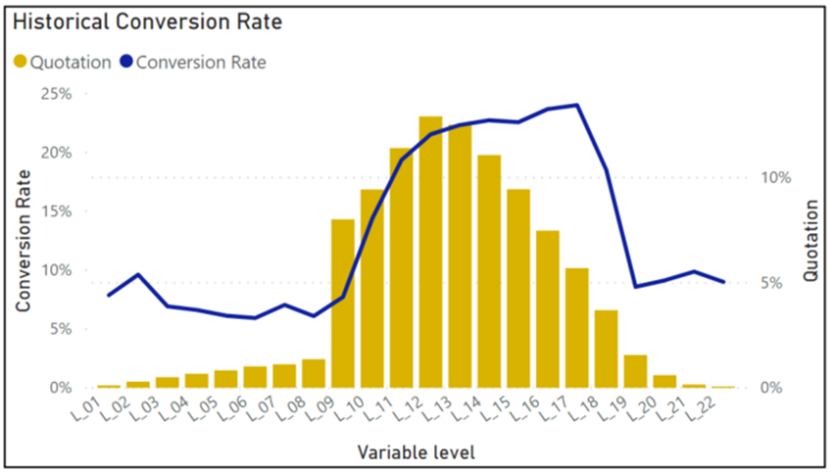

Figure 4.4: V_{03} Conversion Rate graph

Conversion rate: the conversion rates for levels between L_{01} and L_{06} are below average, which means that a reduction in overall price for these bands is a good approach in order to capture part of this market.
On the other hand, the model suggests a reduction in the relativities for levels over L_{14}. This reduction will lead to the maintenance of the average premium for levels between L_{14} and L_{17}, which is desired since our conversion rates are at a good level for this exposure interval.
The next three graphs are related to the Impact Analysis for variable V_{18}:

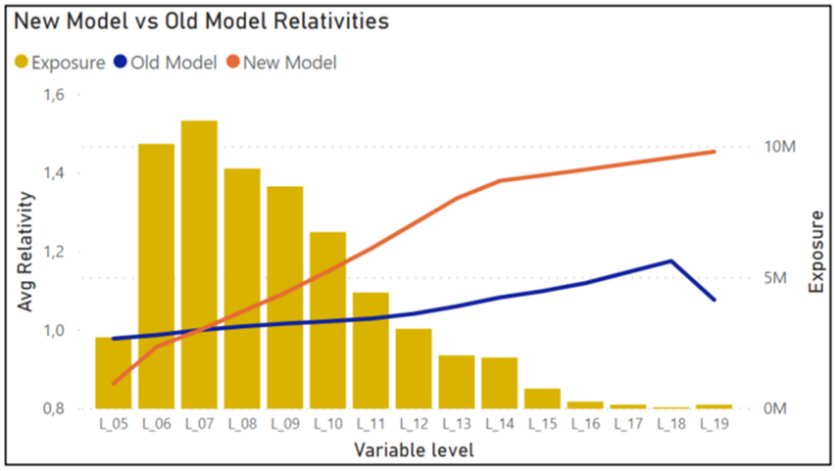

Figure 4.5: V_{18} Relativities graph

Relativities: the new model suggests a small decrease of the relativities in the first two levels, impacting premiums negatively. For the other levels, the suggestion is to increase the relativities, with a bigger impact in the levels over L_{14}.

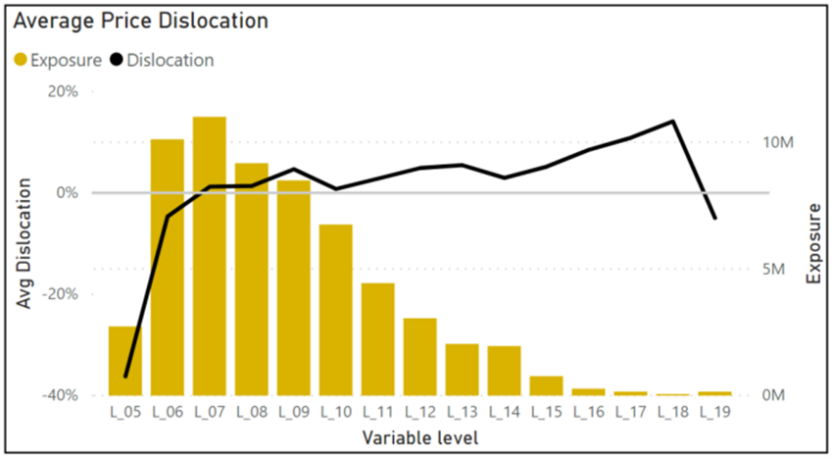

Figure 4.6: V_{18} Average Price Dislocation graph

Price Dislocation: the relativity changes go in the same direction as price dislocation, which means that other variables ratifies the suggested impact on premiums in the same way (with exception of the last level, which has very low exposure). So we can say that, in general, as we increase the relativities for most of the levels in this variable, we will also increase the average price for these same levels.

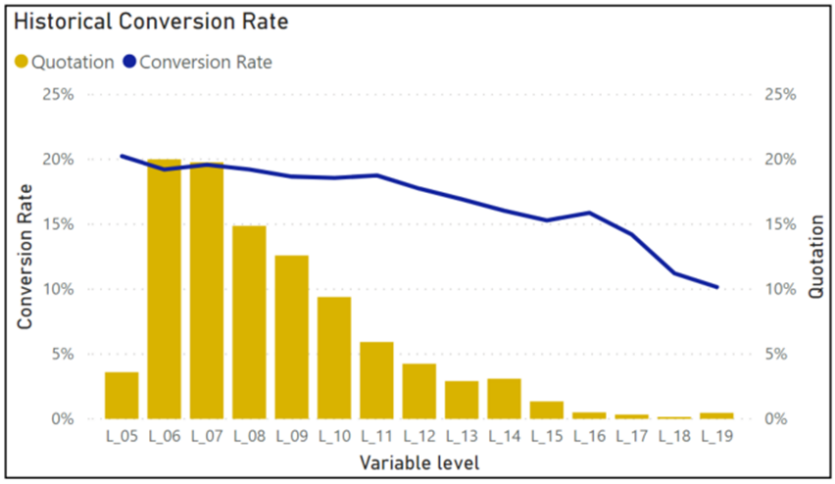

Figure 4.7: V_{18} Conversion Rate graph

Conversion rate: The changes in the premium will have two different effects in conversion. The decrease of 35% in the premium for L_{05} will have a positive impact in conversion, bringing conversion in this level for an even higher value. At the same time, the other levels will be impacted negatively in terms of conversion, which can lead to a loss in market share for these levels of lower exposure.
All this analysis is discussed with the Product Team so we can decide which approach to take when defining the final relativities before launching the product.

Chapter 5

Conclusion

5.1 Results

From holdout validation, we concluded that all variables are significant for all their levels in the models built, which suggests that the models built are consistent with a dataset not used in the training step.
From model evaluation, we concluded that the model has a good response when compared to the new dataset, by analyzing the single lift charts. Also, it was possible to verify the similarity between the new model and the previous one, in terms of the ability to predict the behavior of a new dataset, and both models show a significant improvement over the random selection, as expected.
From the impact analysis, we concluded that the variables have similar trends, by analyzing the relativities, which is a sign of consistency. The combination of relativities, average price dislocation and conversion graphs analysis can lead to insightful suggestions about changes in price for specific categories. All these inputs are of great help when discussing with the Product team and defining the final tariffs that will reach the market.

5.2 Next Steps

This work is limited to show how risk models based on GLM are built. In future researches, it could include the comparison with other methodologies like Machine Learning models. Regarding impact analysis, it could also considers price elasticity as a complement to evaluate the changes in the prices.

Bibliography

[1] Liberty Seguros, '*Liberty Sobre Rodas - Condições Gerais e Especiais*'.
URL: https://www.libertyseguros.pt/Formulario/Documentacao/e303649a-53fc-4d1c-86ae-275e31fd1776

[2] McCullagh, P., Nelder and J. A. [1989], '*Generalized Linear Models*'.
URL: http://www.utstat.toronto.edu/~brunner/oldclass/2201s11/readings/glmbook.pdf

[3] Lindsey, J. K. [1997], '*Applying Generalized Linear Models*'.
URL: http://www.leg.ufpr.br/lib/exe/fetch.php/wiki:internas:biblioteca:lindsey-glm.pdf

[4] Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E. and Thandi, N. [2007], '*A Practitioner's Guide to Generalized Linear Models*'.
URL: https://www.casact.org/pubs/dpp/dpp04/04dpp1.pdf

[5] de Jong, P. and Heller, G. Z. [2008], '*GENERALIZED LINEAR MODELS FOR INSURANCE DATA*'.
URL: https://feb.kuleuven.be/public/u0017833/boek.pdf

[6] Dobson, A. J. and Barnett, A. G. [2008], '*An Introduction to Generalized Linear Models*'.
URL: http://www.ru.ac.bd/wp-content/uploads/sites/25/2019/03/202_06_Dobson_

An-Introduction-to-Generalized-Linear-Models-2008.pdf

[7] Kaas, R., Goovaerts, M., Dhaene, J. and Denuit, M. [2009], '*Modern Actuarial Risk Theory*'.
URL: https://faculty.ksu.edu.sa/sites/default/files/modern_acturial_risk_theory.pdf

[8] Goldburd, M., Khare, A. and Tevet, D. [2019], '*GENERALIZED LINEAR MODELS FOR INSURANCE RATING*'.
URL: https://www.casact.org/sites/default/files/2021-01/05-Goldburd-Khare-Tevet.pdf

[9] Autoridade de Supervisão de Seguros e Fundos de Pensões, '*Produção Provisória - Automóvel (em Portugal)*'.
URL: https://www.asf.com.pt/ISP/Estatisticas/seguros/estatisticas_anuais/premios/ranking_actividade/automovel.htm

Appendix A

Glossary

AICc: Akaike Information Criterion corrected for small sample sizes
ASF: Autoridade de Supervisão de Seguros e Fundos de Pensões
B-M: Bonus-Malus
CCC: *Choque, Colisão e Capotamento*, which means Shock, Collision and Rollover
CF: Custom Factor type of fit
CL: Variable related to claims history information
DR: Variable related to driver information
GEO: Variable related to geographical information
GLM: Generalized Linear Model
RC: *Responsabilidade Civil*, which means Third Party Liability
RCC: *Responsabilidade Civil Corporal*, which means Third Party Liability - Bodily Injury
RCM: *Responsabilidade Civil Material*, which means Third Party Liability - Property Damage
SA: Variable related to policy information
SF: Simple Factor type of fit
TI: Variable related to time information
TPL: Third Party Liability
VH: Variable related to vehicle information
VR: Variate type of fit

Appendix B

Variables Tables

Variable	Category	Fit Type
V_{01}	Time Information (TI)	Simple Factor (SF)
V_{02}	Driver Information (DR)	Simple Factor (SF)
V_{03}	Policy Information (SA)	Variate (VR)
V_{04}	Policy Information (SA)	Simple Factor (SF)
V_{05}	Policy Information (SA)	
V_{06}	Policy Information (SA)	Simple Factor (SF)
V_{07}	Policy Information (SA)	Variate (VR)
V_{08}	Policy Information (SA)	Custom Factor (CF)
V_{09}	Policy Information (SA)	
V_{10}	Vehicle Information (VH)	Variate (VR)
V_{11}	Vehicle Information (VH)	Custom Factor (CF)
V_{12}	Vehicle Information (VH)	
V_{13}	Vehicle Information (VH)	Custom Factor (CF)
V_{14}	Vehicle Information (VH)	Custom Factor (CF)
V_{15}	Vehicle Information (VH)	Custom Factor (CF)
V_{16}	Vehicle Information (VH)	Variate (VR)
V_{17}	Vehicle Information (VH)	Variate (VR)
V_{18}	Vehicle Information (VH)	Variate (VR)
V_{19}	Vehicle Information (VH)	Variate (VR)
V_{20}	Geographical Information (GEO)	Custom Factor (CF)
V_{21}	Geographical Information (GEO)	Custom Factor (CF)
V_{22}	Claims History (CL)	Simple Factor (SF)

Table B.1: List of Variables

	Name	Value	Standard Error (%)	Weight (%)	Exp(Value)
01	Mean	-3.244	0.5	100	0.0390
-	$V_{02} - L_{01}$			65.3[1]	
02	$V_{02} - L_{02}$	0.1003	11.8	22.9	1.1055
03	$V_{02} - L_{03}$	0.1724	11.3	9.5	1.1882
04	$V_{02} - L_{04}$	0.2071	14.1	2.3	1.2302
05	$V_{04} - L_{01}$	-0.0743	15.8	23.6	0.9284
-	$V_{04} - L_{02}$			69.6	
06	$V_{04} - L_{03}$	-0.2716	8.6	6.8	0.7622
-	$V_{06} - L_{01}$			66.7	
07	$V_{06} - L_{02}$	0.2756	4.6	16.0	1.3174
08	$V_{06} - L_{03}$	0.4870	3.3	7.0	1.6274
09	$V_{06} - L_{04}$	0.2585	6.2	10.2	1.2950
-	$V_{22} - L_{01}$			49.7	
10	$V_{22} - L_{02}$	0.1083	14.0	17.5	1.1143
11	$V_{22} - L_{03}$	0.2751	8.5	3.8	1.3166
12	$V_{22} - L_{04}$	0.3960	5.9	3.4	1.4858
13	$V_{22} - L_{05}$	0.3594	7.8	2.4	1.4325
14	$V_{22} - L_{06}$	0.2860	11.0	2.0	1.3311
15	$V_{22} - L_{07}$	0.2755	11.1	2.2	1.3172
16	$V_{22} - L_{08}$	0.2260	13.9	2.2	1.2535
17	$V_{22} - L_{09}$	0.2710	11.4	2.2	1.3113
18	$V_{22} - L_{10}$	0.3255	7.5	4.0	1.3847
19	$V_{22} - L_{11}$	0.4626	14.3	0.3	1.5882
20	$V_{22} - L_{12}$	0.6775	9.2	0.3	1.9689
21	$V_{22} - L_{13}$	0.7486	9.7	0.2	2.1140
22	$V_{22} - L_{14}$	0.8279	11.7	0.1	2.2885
23	$V_{22} - L_{15}$	0.7058	16.0	0.1	2.0256
24	$V_{22} - L_{16}$	0.7512	16.2	0.1	2.1196
25	$V_{22} - L_{17}$	0.8480	14.8	0.0	2.3349
26	$V_{22} - L_{18}$	0.9124	10.5	0.1	2.4903
27	$V_{22} - L_{19}$	0.1008	17.6	9.5	1.1060

[1]The blue values represent the base level for the corresponding variable

Continuation of Table 4.9

	Name	Value	Standard Error (%)	Weight (%)	Exp(Value)
28	$V_{14} - CF_{01}$	0.0213	10.0	11.8	1.2230
29	$V_{14} - CF_{02}$	0.1146	14.1	66.0	1.1214
-	$V_{14} - CF_{03}$			22.2	
-	$V_{15} - CF_{01}$			86.3	
30	$V_{15} - CF_{02}$	0.1003	27.0	0.3	0.7168
31	$V_{15} - CF_{03}$	0.1724	8.5	13.3	1.2351
32	$V_{20} - CF_{01}$	-0.2049	12.9	6.5	0.8147
33	$V_{20} - CF_{02}$	-0.1672	11.4	31.2	0.8461
-	$V_{20} - CF_{03}$			45.2	
34	$V_{20} - CF_{04}$	0.1146	10.7	17.0	1.2081
35	$V_{21} - CF_{01}$	-0.5036	10.7	1.1	0.0643
36	$V_{21} - CF_{02}$	-0.2957	7.7	10.9	0.7440
37	$V_{21} - CF_{03}$	-0.2905	7.7	7.3	0.7479
38	$V_{21} - CF_{04}$	-0.1858	10.2	48.2	0.8304
-	$V_{21} - CF_{05}$			32.5	
39	$V_{03} - VR_{01}$	0.0751	7.1	100	1.0779
40	$V_{07} - VR_{01}$	-0.1395	6.4	100	0.8698
41	$V_{10} - VR_{01}$	0.1706	6.7	100	1.1860
42	$V_{10} - VR_{02}$	-0.0652	11.1	100	0.9368
43	$V_{17} - VR_{01}$	-0.1884	6.0	100	0.8283
44	$V_{17} - VR_{02}$	-0.2532	4.5	100	0.7763
45	$V_{16} - VR_{01}$	0.0818	16.4	100	1.0852
46	$V_{16} - VR_{02}$	-0.0530	14.1	100	0.9484

Table B.2: RCM-F fitted variables and statistics