

Health Insurance Cross Sell Prediction

Eva Salivonik

Tuesday, October 18th 2023

- Business Introduction
 - Abstract
 - Background
 - Objective
 - Business Understanding:
- Data Understanding:
 - Data Description :
 - Loading Data
 - View the contents and structure of our dataset:
- View the main structure of the raw data
 - Helper Function
- Exploratory Data Analysis (EDA)
 - What affect customer response?
 - Response by previous insurance
 - Response by past vehicle damage
 - Response by gender
 - Response by vehicle age
 - Response by age
 - Response by Annual Premium
 - What is affecting vehicle damage?
 - Age by Vehicle Damage
 - Vehicle damage by vehicle age
 - vehicle damage by gender
 - Annual Premium
 - annual premium by vehicle damage
 - Age distribution by Vehicle Age
 - Policy sales channel
- Correlation
 - Loading head of training data
 - How we proceed:
 - We will tune some XGB parameters with gridsearch.
 - Oversampling for gridsearch
 - XGB GRID SEARCH
- XGB model
 - Launch XGBoost
- Test XGB on a balanced dataset
 - Rebalance dataset
 - Launch Xgboost
- Submission (Checking Results)
 - Submit XGB predictions
 - Model Deployment
- Conclusion

Business Introduction

Abstract

Cross-selling is a sales strategy employed to encourage customers to increase their spending by acquiring a product that complements their current purchase. This approach finds utility in the insurance industry, where it serves as a means for companies to introduce new products to their existing customer base. Leveraging machine learning models can streamline the labor-intensive process of sifting through customer records, resulting in significant time and cost savings. However, the implementation of machine learning models poses its own set of challenges, some of which we aim to address in this project. Recognizing the pivotal role of data in business, we prioritize customer privacy, adhering to the principles of the Ethical Machine Learning framework. Our team conducts data analysis and model development following the CRISP-DM methodology.

Background

Triks Insurance Inc., a prominent Life Insurance Agent in the town of XYZ, boasts an extensive client base of 381,109 households within the region. Presently, they are in negotiations with a prominent Auto Insurance provider, GOOL Auto Insurance, with plans to introduce their insurance product. GOOL Auto Insurance has requested Triks Insurance to furnish a report indicating the proportion of their client base likely to express interest in Auto insurance.

Rather than embarking on the arduous task of manually sifting through their customer database, Triks Insurance Inc. has opted to engage the SK team to create a machine learning system. This system's purpose is not only to predict potential Auto Insurance candidates once but continuously over time. By comprehending the customer base and leveraging demographic information about a prospect, this machine learning system will assess whether a prospect falls into the category of a "good" or "bad" candidate.

By implementing this solution, Triks Insurance's sales team will be better prepared for their cross-selling endeavors. They can hone their focus on the most promising leads, thus maximizing the efficiency of their marketing budget while concentrating on the most lucrative prospects.

The proposed solution serves several valuable purposes:

- *It provides GOOL Auto with insights into the potential business that can be generated from Triks Insurance's existing client base.
- *It equips the staff at Triks Insurance with a tool for client prioritization, facilitating focused and effective marketing campaigns.
- *It enables Triks Insurance to project future revenues, enhancing their financial planning.
- *The system offers real-time notifications for potential cross-selling opportunities.

The client was initially unaware that the data they possess holds the answers to their questions. Some of their concerns and inquiries include:

- What information or resources do you require from us to deliver this solution?
- Can the system accurately classify customers?
- How will the system handle cases where a customer who should have qualified for GOOL Insurance is not selected?

Objective

The aim of this project is to create one or more models to generate an initial report for GOOL Insurance on behalf of Triks Insurance Ltd. Additionally, the project seeks to develop a predictive tool that can determine whether a customer, based on specific characteristics, qualifies as a potential candidate for their upcoming cross-selling campaign.

Business Understanding:

To understand and address business problems:

- Determine the key variables to predict (good/bad candidates). -Define relevant metrics related to these variables.
- Grasp project objectives and requirements. -Create specific questions, such as what distinguishes a good from a bad prospect. -Assess current sales practices for identifying prospects. -Establish a success metric for the project.
- Identify data sources with answers to these questions. -Select data that accurately measures the target and relevant features. -Consider if the existing system requires additional data for the project. -Evaluate the need for external data sources or system updates.

Then, translate this knowledge into a data mining problem and create a preliminary plan to achieve the project goals.

Data Understanding:

Having understood the business statement, it's clear that Triks Insurance faced a choice. They could have promoted Auto Insurance to all their customers, but that wouldn't have been the most efficient use of their marketing budget. It's more cost-effective to focus on customers likely to respond to the GOOL Auto campaign. This targeted approach saves money and doesn't bother customers uninterested in the new product.

To achieve this, we can leverage historical data from past campaigns to build a model for predicting which customers are promising prospects. We start by collecting relevant data, addressing data quality issues, gaining initial insights from the data, and identifying interesting trends or patterns within it.

The data for this project was downloaded from Kaggle, weblink: <https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction> (<https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction>). For privacy and security, the customer names have been masked with an id. The data is in csv format.

Data Description :

Variable Name	Variable Description
id	Unique ID for the customer
Gender	Gender of the customer
Age	Age of the customer
Driving_License	0 : Customer does not have DL 1 : Customer already has DL
Region_Code	Unique code for the region of the customer
Previously_Insured	1 : Customer already has Vehicle Insurance.

Variable Name	Variable Description
	0 : Customer doesn't have Vehicle Insurance.
Vehicle_Age	Age of the Vehicle
Vehicle_Damage	1 : Customer's vehicle damaged in the past.
	0 : Customer's vehicle no damaged in the past.
Annual_Premium	The premium insured paid in the year
Policy_Sales_Channel	Anonymized Code for outreach channels for sales.
Vintage	Number of Days, Customer has been associated with.
	the company
Response	1 : Customer is interested
	0 : Customer is not interested

Loading Data

```
test <- read.csv("c://DATA//test.csv")
train <- read.csv('c://DATA//train.csv')
submission <- read.csv('c://DATA//sample_submission.csv')
```


###Determine the dimension of our dataset:

```
## [1] 381109    12
```

View the contents and structure of our dataset:

	id	Gen...	A..	Driving_License	Region_Co...	Previously_Insured	Vehicle_Age	Vehicle_Damr
	<int>	<chr>	<int>	<int>	<dbl>	<int>	<chr>	<chr>
1	1	Male	44	1	28	0	> 2 Years	Yes
2	2	Male	76	1	3	0	1-2 Year	No
3	3	Male	47	1	28	0	> 2 Years	Yes
4	4	Male	21	1	11	1	< 1 Year	No
5	5	Female	29	1	41	1	< 1 Year	No
6	6	Female	24	1	33	0	< 1 Year	Yes

6 rows | 1-9 of 13 columns



id	Gen...	A..	Driving_License	Region_Co...	Previously_Insured	Vehicle_Age	Vehicle_
<int>	<chr>	<int>	<int>	<dbl>	<int>	<chr>	<chr>
1 381110	Male	25	1	11	1	< 1 Year	No
2 381111	Male	40	1	28	0	1-2 Year	Yes
3 381112	Male	47	1	28	0	1-2 Year	Yes
4 381113	Male	24	1	27	1	< 1 Year	Yes
5 381114	Male	27	1	28	1	< 1 Year	No
6 381115	Male	22	1	30	1	< 1 Year	No

6 rows | 1-9 of 12 columns

View the main structure of the raw data

```
## 'data.frame': 381109 obs. of 12 variables:
## $ id : int 1 2 3 4 5 6 7 8 9 10 ...
## $ Gender : chr "Male" "Male" "Male" "Male" ...
## $ Age : int 44 76 47 21 29 24 23 56 24 32 ...
## $ Driving_License : int 1 1 1 1 1 1 1 1 1 1 ...
## $ Region_Code : num 28 3 28 11 41 33 11 28 3 6 ...
## $ Previously_Insured : int 0 0 0 1 1 0 0 0 1 1 ...
## $ Vehicle_Age : chr "> 2 Years" "1-2 Year" "> 2 Years" "< 1 Year" ...
## $ Vehicle_Damage : chr "Yes" "No" "Yes" "No" ...
## $ Annual_Premium : num 40454 33536 38294 28619 27496 ...
## $ Policy_Sales_Channel: num 26 26 26 152 152 160 152 26 152 152 ...
## $ Vintage : int 217 183 27 203 39 176 249 72 28 80 ...
## $ Response : int 1 0 1 0 0 0 0 1 0 0 ...
```

Helper Function

Helper functions are used in conjunction with the main verbs to make specific tasks and calculations a bit easier.

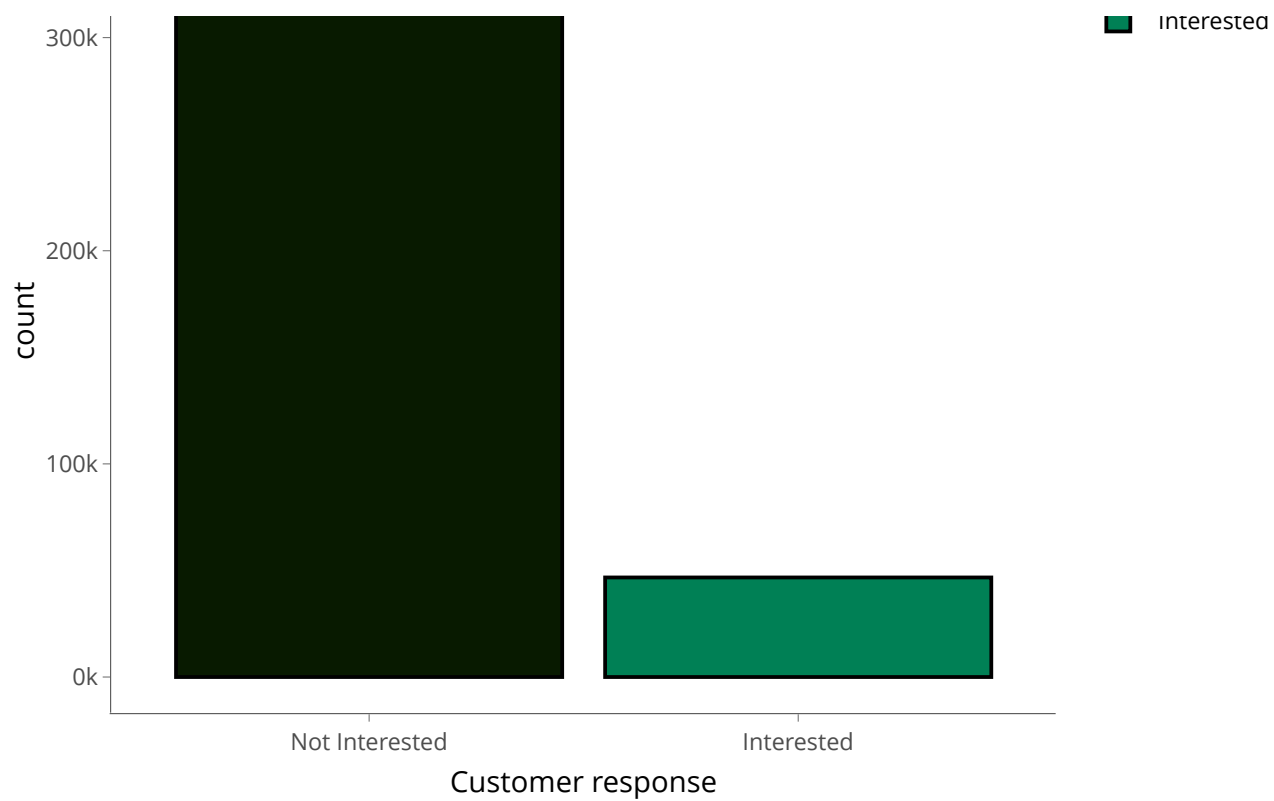
Exploratory Data Analysis (EDA)

We can create a profile of the ideal customer for our insurance company:

The ideal customer doesn't currently have active insurance. They have a history of vehicle accidents or damage. Their vehicle is relatively new, between 1 to 2 years old. Their age range of 30 to 55. Their annual premium between 30,000 and 40,000. They are more inclined to use channels like channel 25 or channel 125, rather than channel 155. ### Customer Response

Customer Response





What affect customer response?

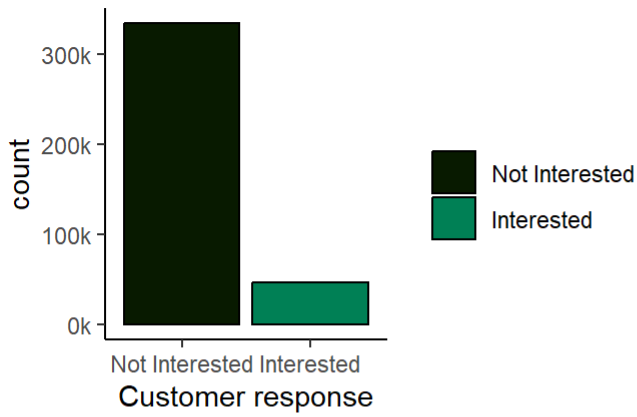
Response by previous insurance

Response by past vehicle damage

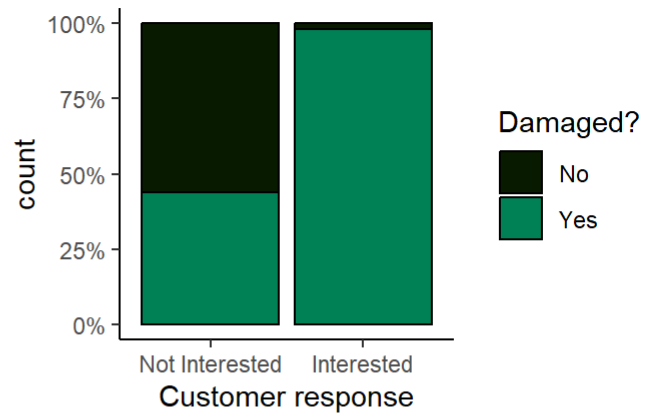
Response by gender

Response by vehicle age

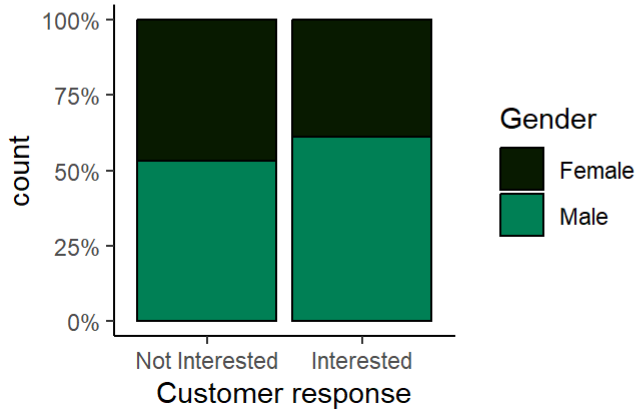
Customer Response



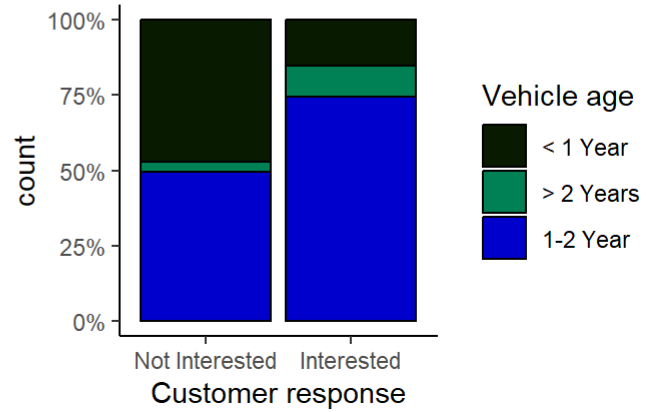
Response by vehicle history



Response by gender

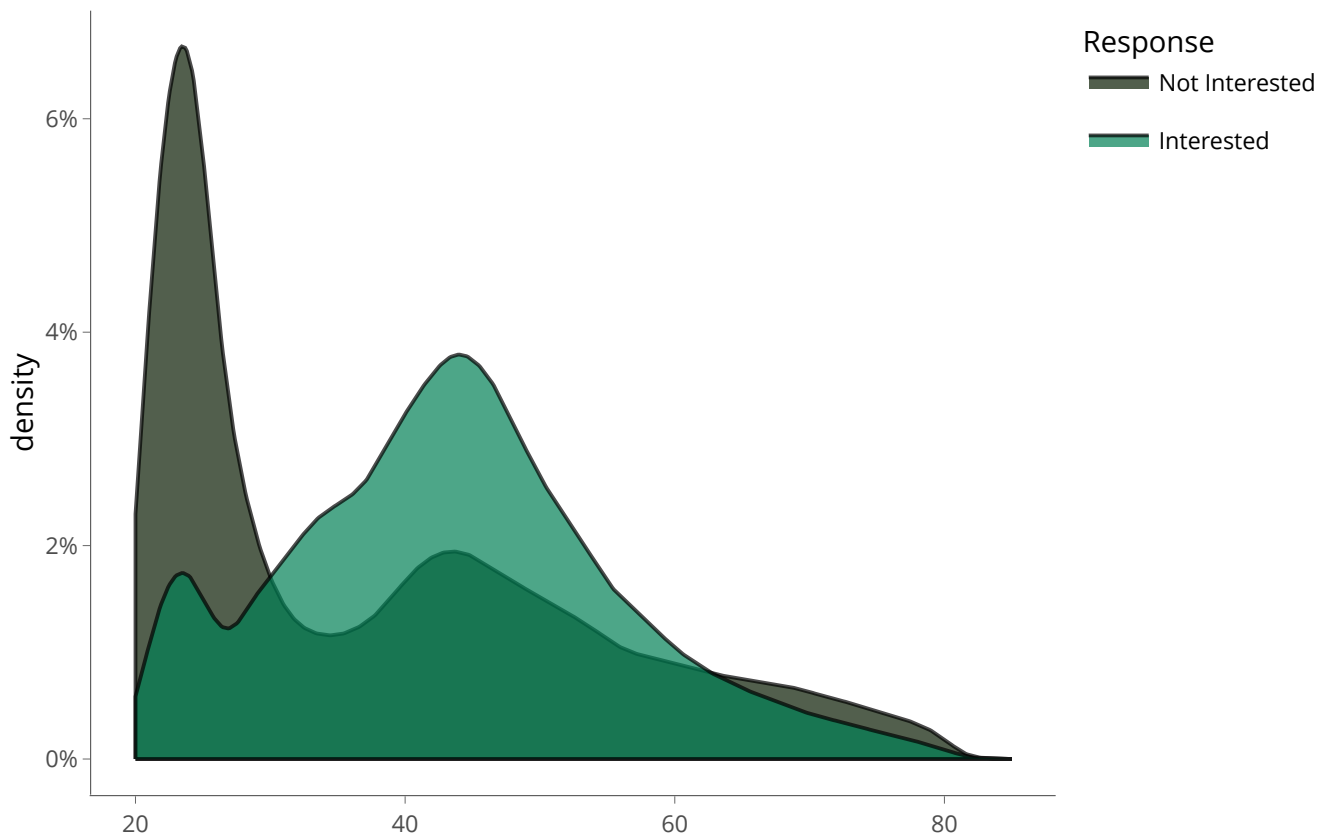


Response by vehicle age



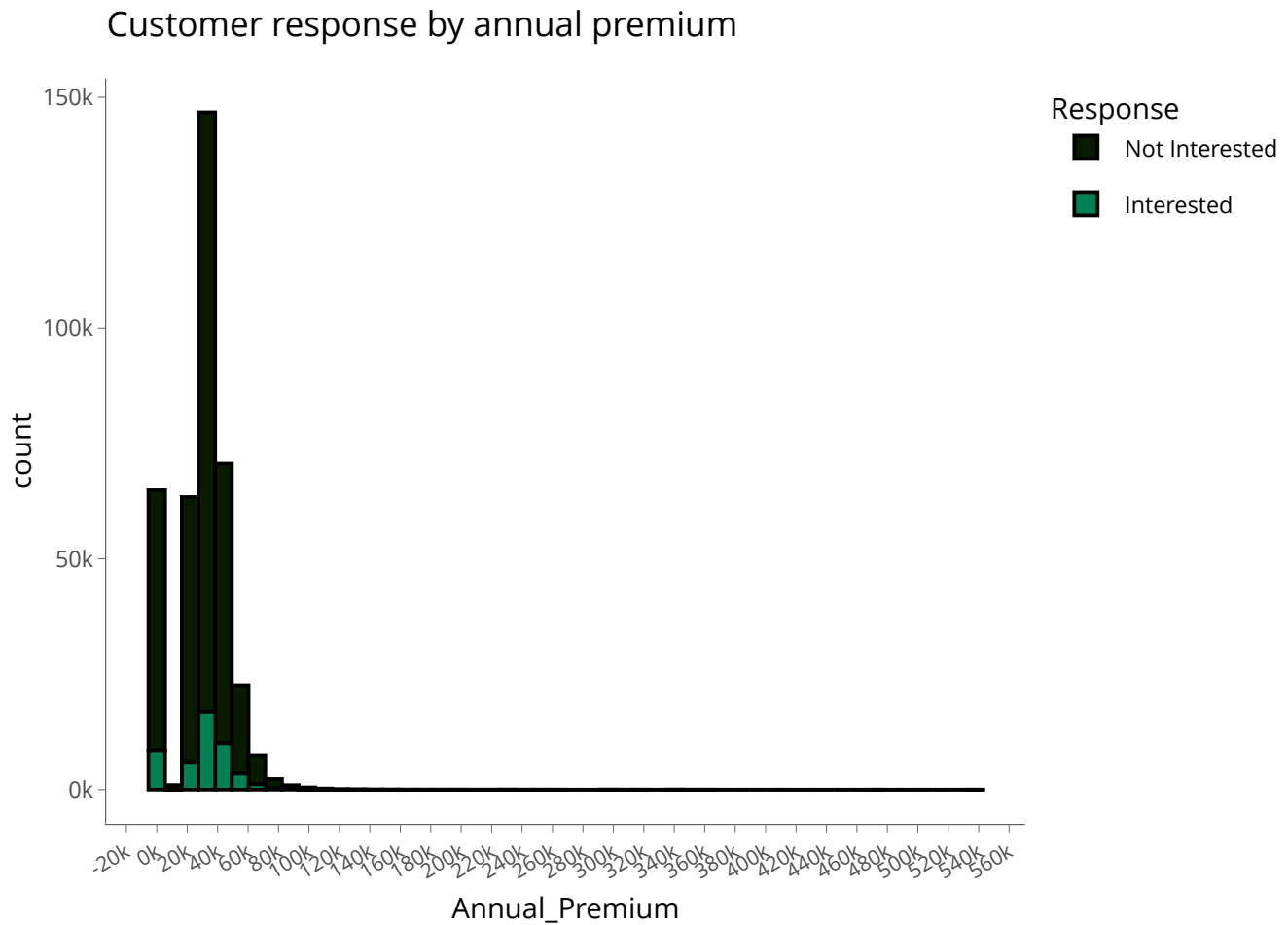
Response by age

Customer response by age



Age

Response by Annual Premium



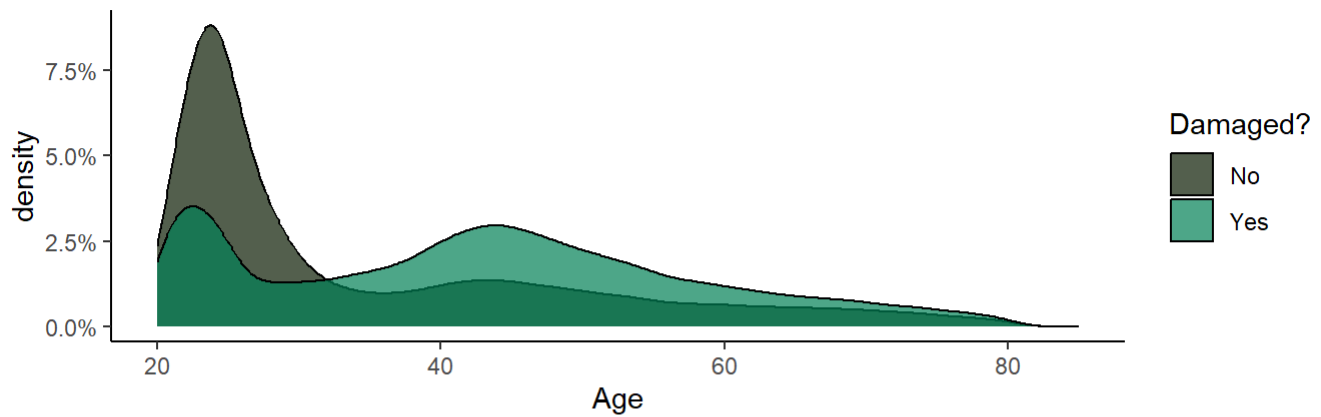
What is affecting vehicle damage?

Age by Vehicle Damage

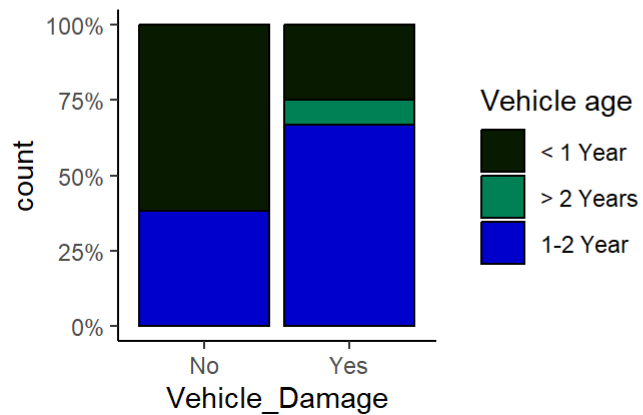
Vehicle damage by vehicle age

vehicle damage by gender

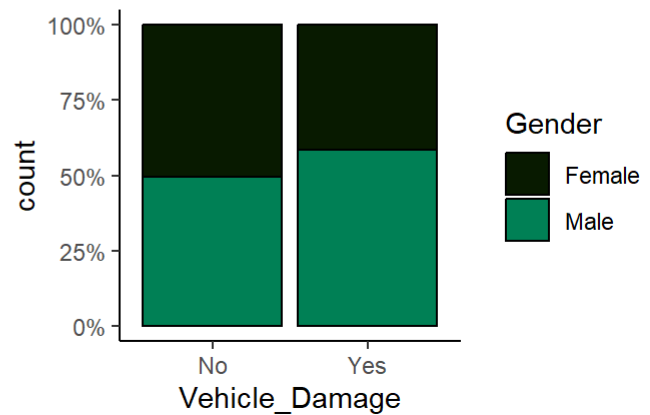
Who had a vehicle damaged?



Vehicle damaged by age



Vehicle damage by gender

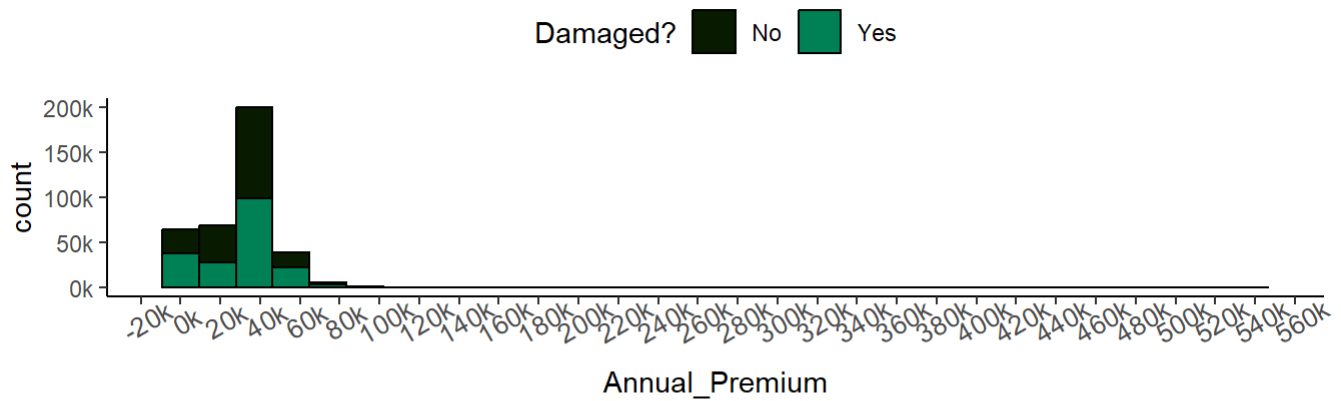


Annual Premium

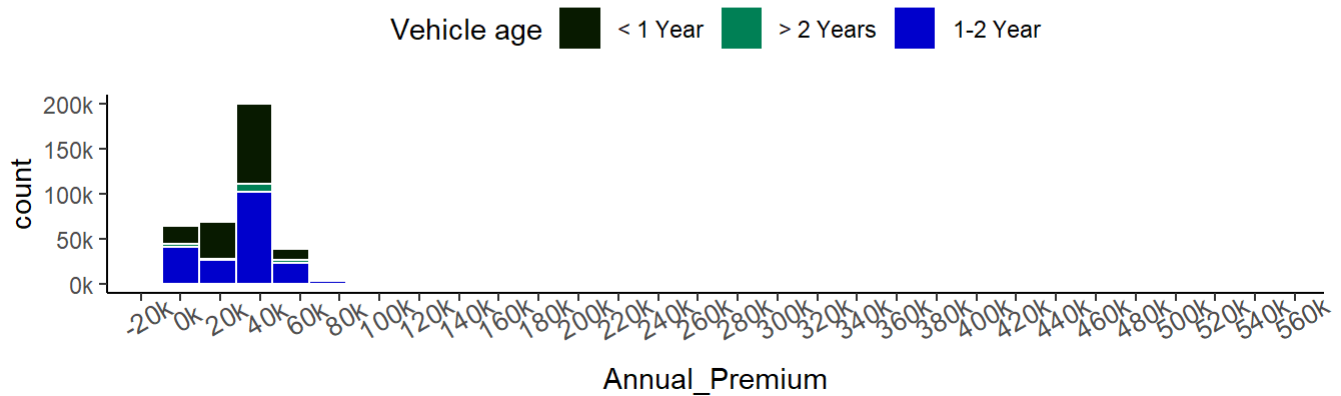
annual premium by vehicle damage

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Annual Premium based on past vehicle damages

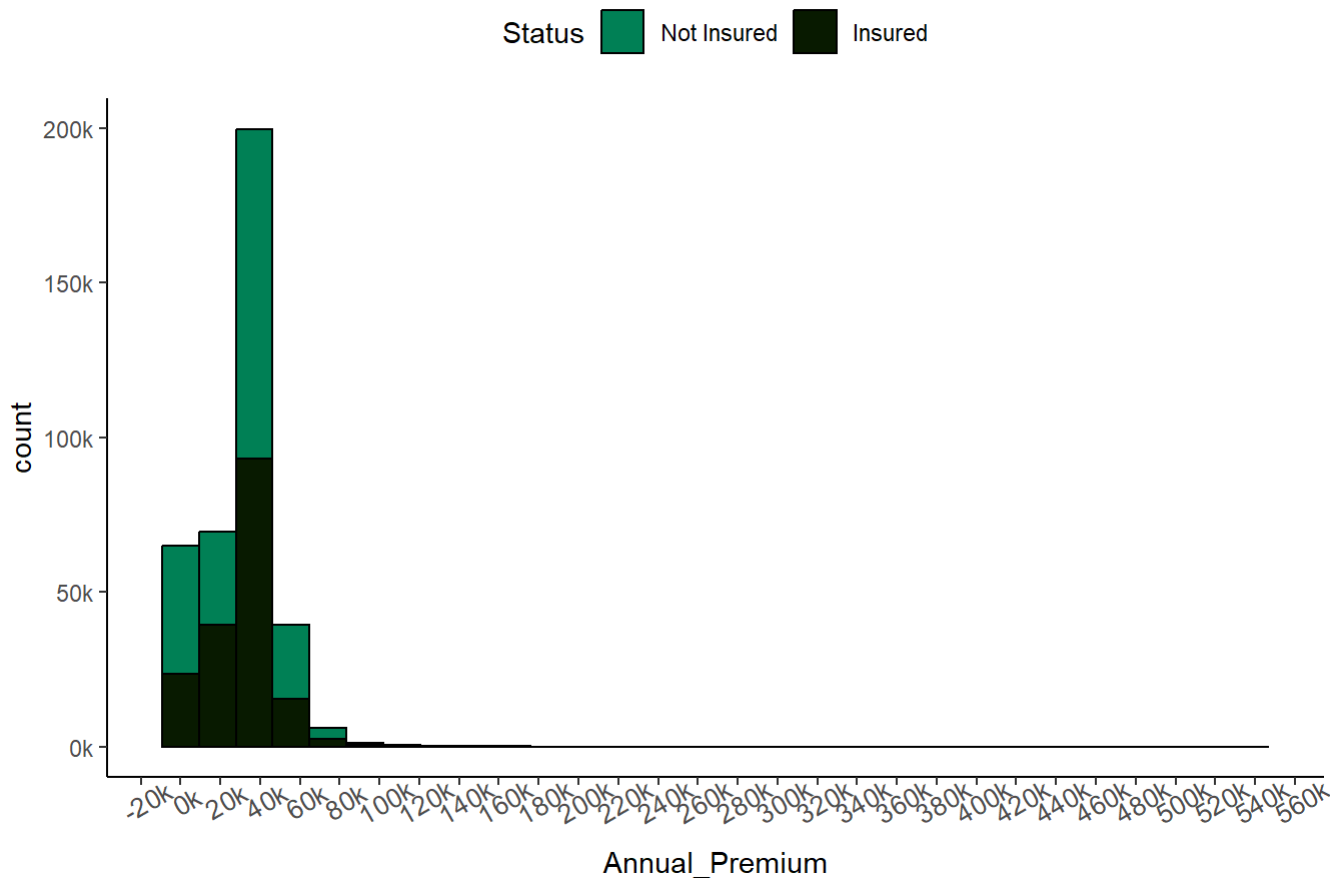


Annual Premium based on Vehicle Age

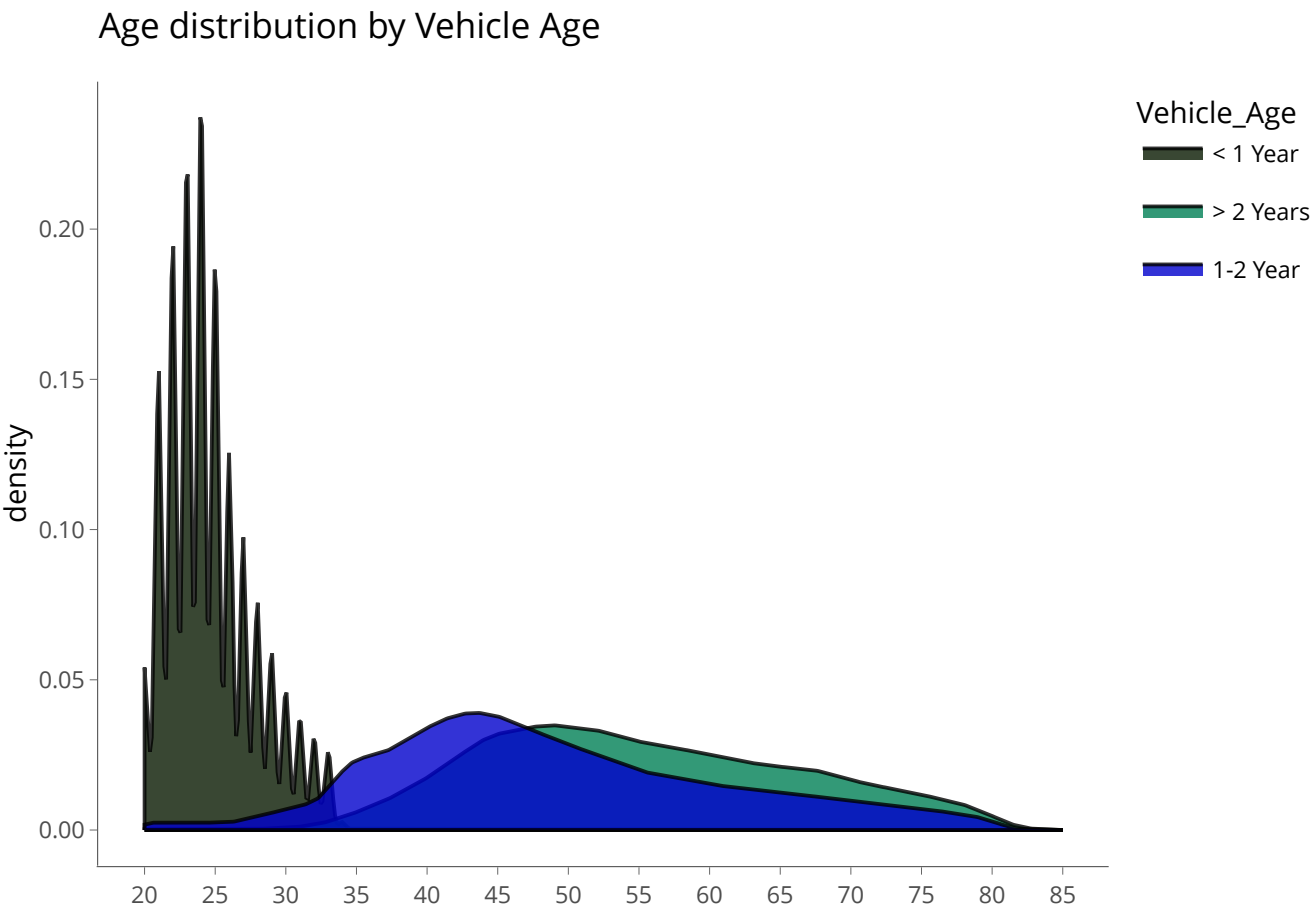


```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Annual Premium based on Vehicle Age



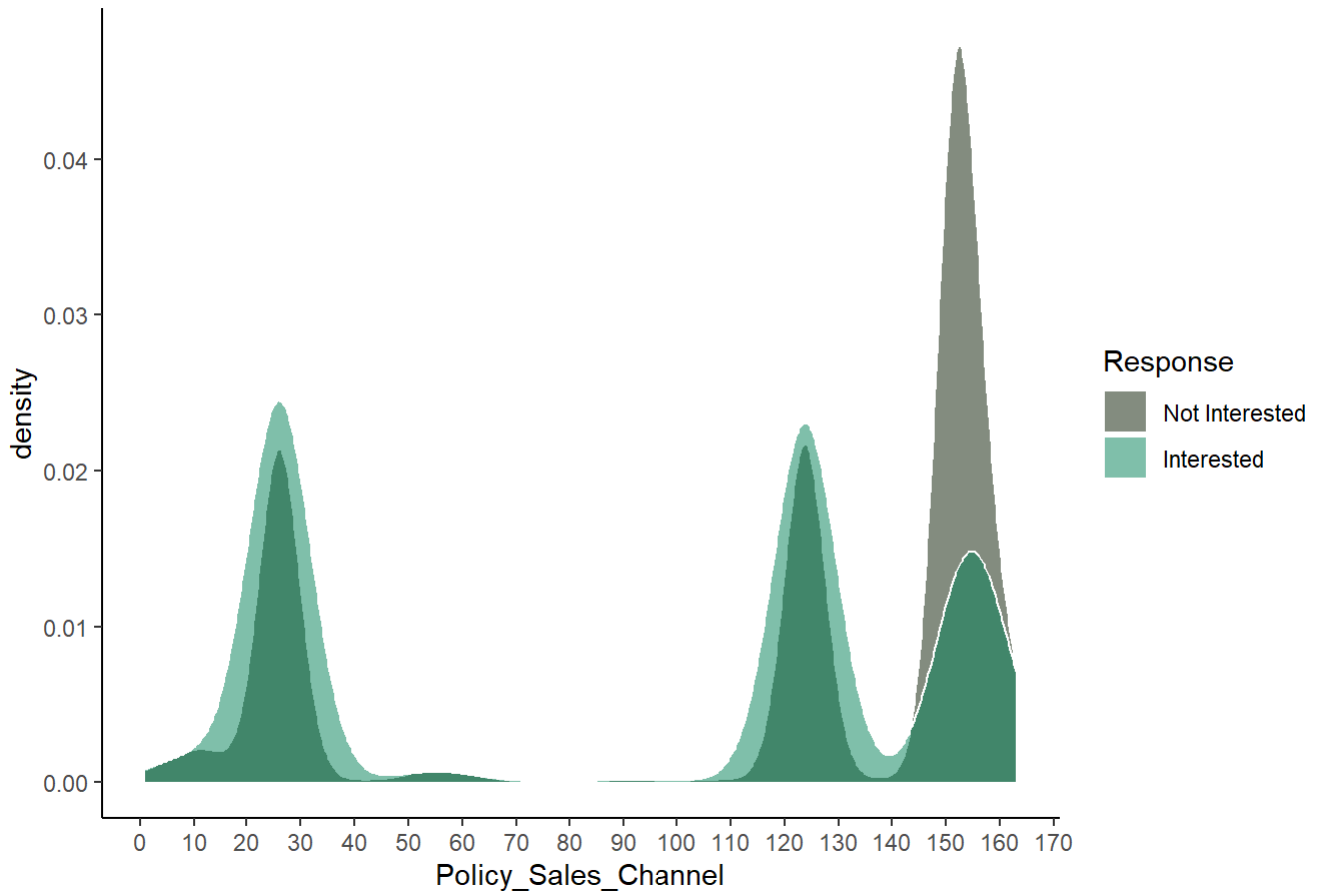
Age distribution by Vehicle Age



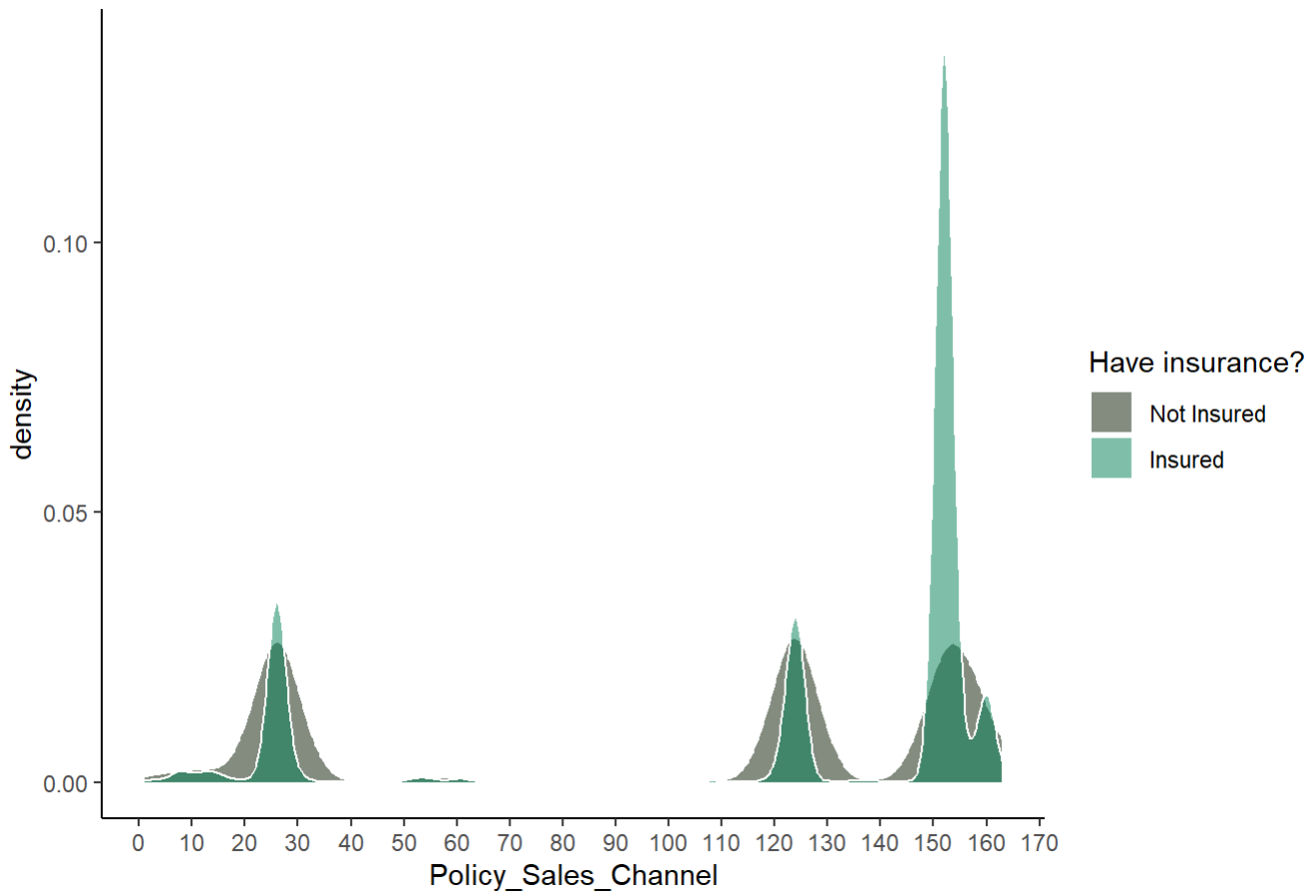
Age

Policy sales channel

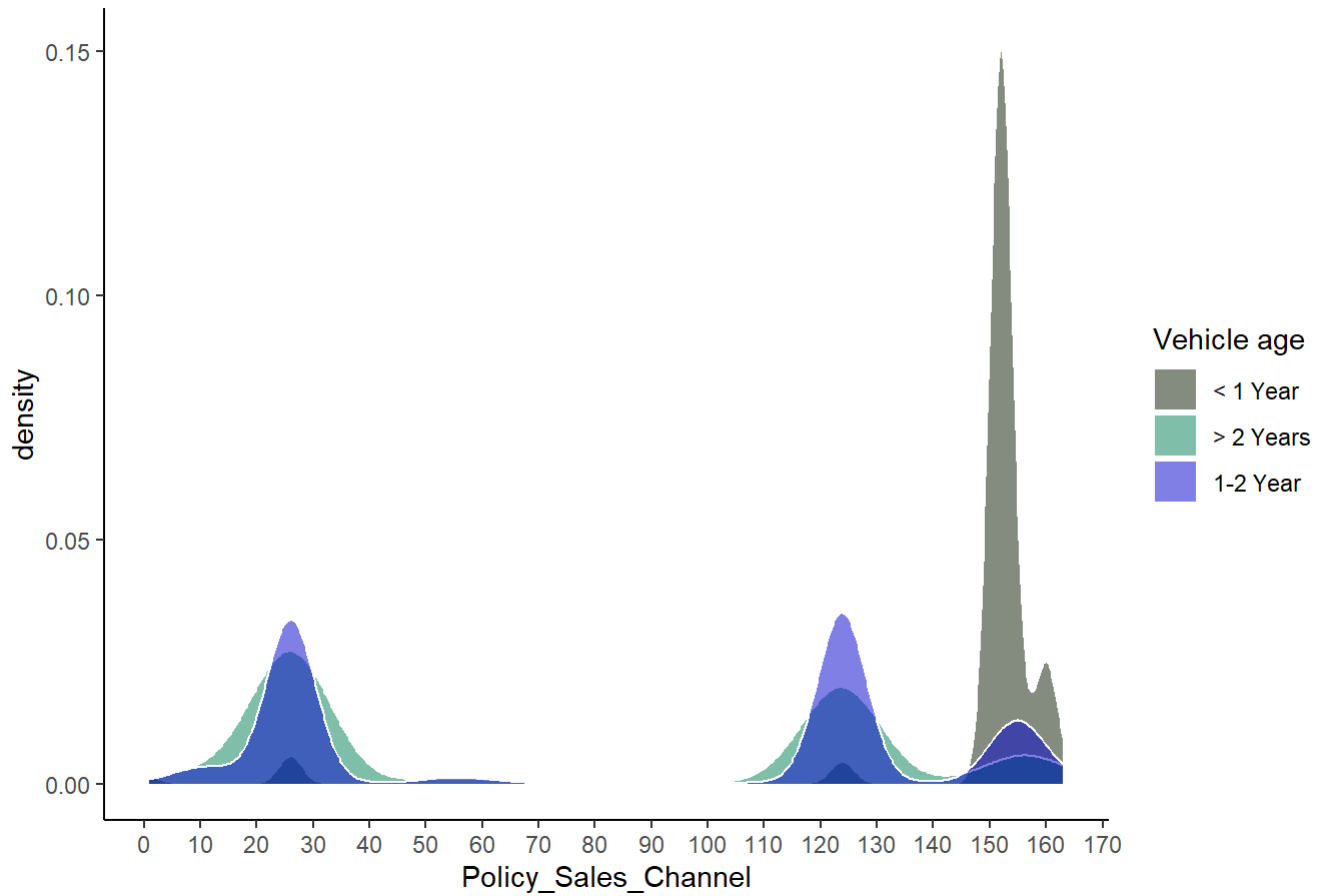
Response by sales channel



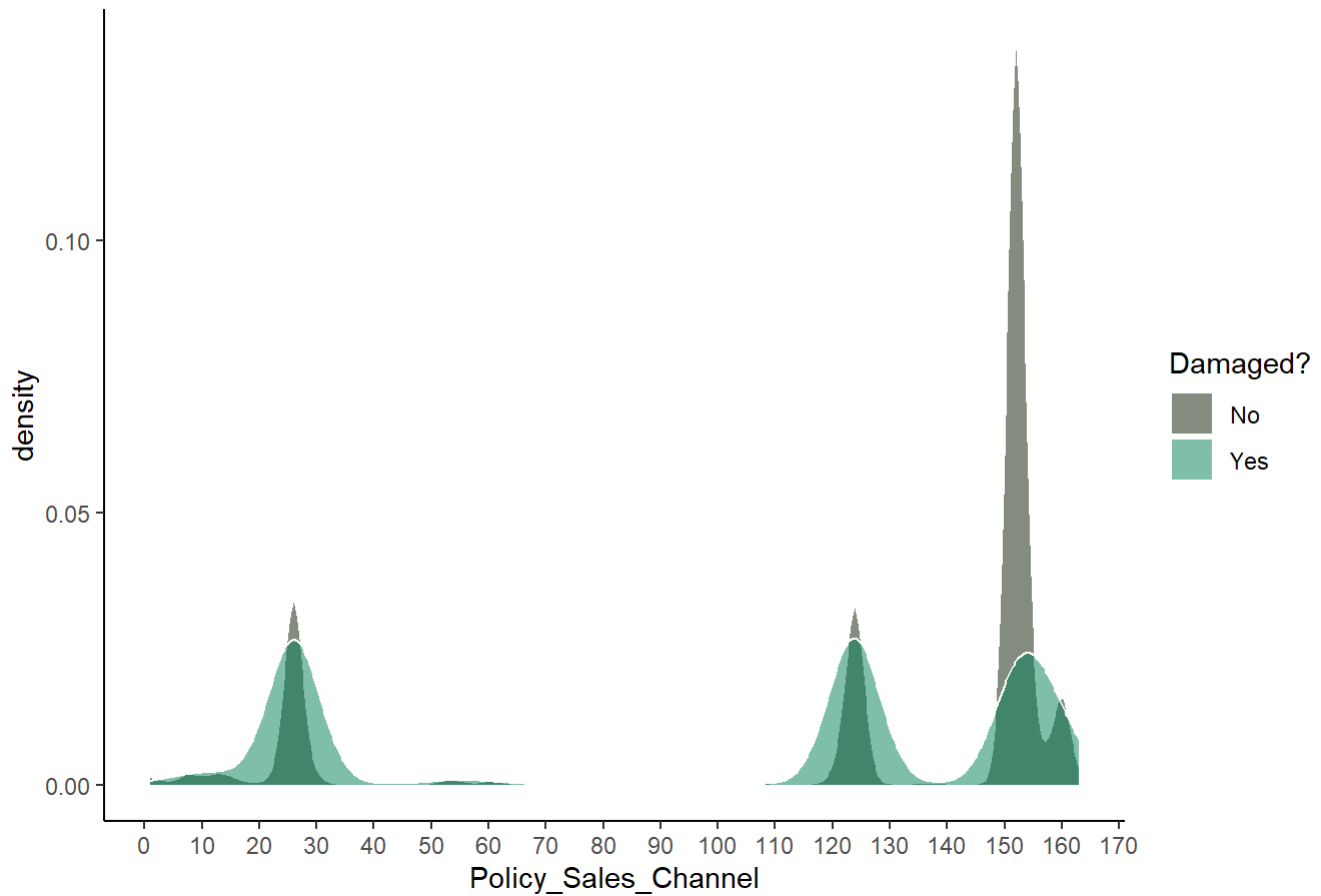
Distribution of sales channel by Insured/Not insured customers



Distribution of sales channel by Vehicle age

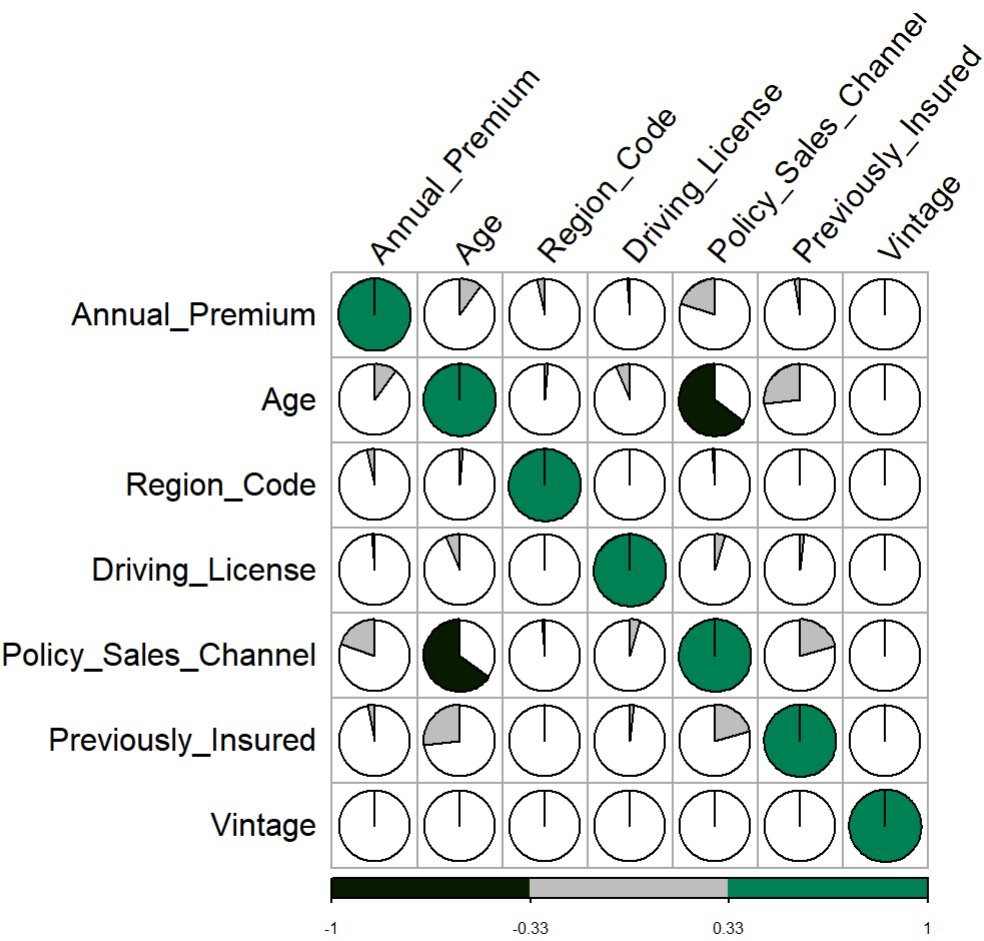


Distribution of sales channel by Vehicle Damage



Correlation

Examine the relationships between independent numerical variables by creating a correlation matrix to assess the correlation coefficients between these variables. Before constructing the correlation matrix, exclude the 'id' and 'Response' columns(correlation can be perform on numeric only).



Loading head of training data

[1] "data.frame"

	Age <dbl[,1]>	Driving_License <dbl[,1]>	Region_Co... <dbl[,1]>	Previously_Insured <dbl[,1]>	Annual_Premium <dbl[,1]>	Policy_S
1	0.3337768	0.04620787	0.1217843	-0.9196368	0.5745379	
2	2.3967476	0.04620787	-1.7678764	-0.9196368	0.1726360	
3	0.5271803	0.04620787	0.1217843	-0.9196368	0.4490525	
4	-1.1489834	0.04620787	-1.1631850	1.0873830	-0.1130176	
5	-0.6332407	0.04620787	1.1044079	1.0873830	-0.1782584	
6	-0.9555799	0.04620787	0.4997164	-0.9196368	-1.6228512	

6 rows | 1-7 of 8 columns

##Create The XGBoost model XGBoost is a boosting technique in machine learning, known for its ability to generate highly accurate predictive models.

####How deal with an imbalanced dataset? There are numerous approaches, and while many of them are valid, this time we decided to take a different path.

How we proceed:

Our approach is as follows:

We'll fine-tune certain XGBoost parameters using grid search. We'll apply the XGBoost Classifier to the imbalanced dataset (the original dataset), while setting the scale_pos_weight : sum(negative)/sum(positive). We'll assess the XGBoost Classifier's performance on a balanced dataset that matches the length of the training dataset. Finally, we'll submit the results.

Since all customers in the submission dataset are uninterested, we will focus solely on evaluating the model's metrics.

We will tune some XGB parameters with gridsearch.

Oversampling for gridsearch

XGB GRID SEARCH

```
## [1] "Best hyperparameters combination: "
```

```
## Warning in train.default(x = as.matrix(grid_train %>% select(-Response)), : The
## metric "Accuracy" was not in the result set. ROC will be used instead.
```

```
##      nrounds max_depth eta gamma colsample_bytree min_child_weight subsample
## 40         10        6 0.3    0              0.3                2        0.5
```

XGB model

####scale_pos_weight= 334399/46710 = 7.15

Response	#
0	334399
1	46710

Launch XGBoost

Test XGB on a balanced dataset

As a further test for our model, we will rebalance the original dataset and evaluate the model performance on a balanced dataset

Rebalance dataset

```
## [1] "Length of train_3 dataset(Balanced dataset): 381109"
```

Response	%
0	0.4997101
1	0.5002899

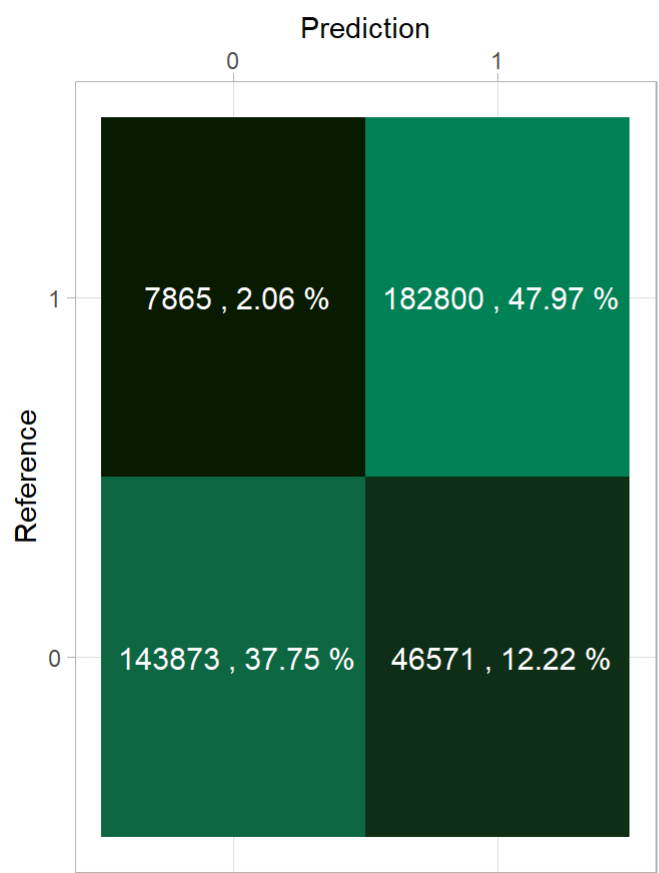
Launch Xgboost

```
## [1] "XGB is starting..."
## [1] "XGB Train:"
## [1]  train-auc:0.798917
## [201]  train-auc:0.883364
## [401]  train-auc:0.899102
## [601]  train-auc:0.910498
## [801]  train-auc:0.921392
## [1000] train-auc:0.930757
## [1] "XGB Cross Validation:"
## [1]  train-auc:0.841687+0.000625 test-auc:0.841301+0.001094
## [201]  train-auc:0.908711+0.000917 test-auc:0.885836+0.001570
## [401]  train-auc:0.937619+0.001331 test-auc:0.904483+0.001312
## [601]  train-auc:0.956170+0.000755 test-auc:0.917808+0.001107
## [801]  train-auc:0.967871+0.000623 test-auc:0.927173+0.001033
## [1000] train-auc:0.976454+0.000681 test-auc:0.934779+0.000742
```

```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
## of ggplot2 3.3.4.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

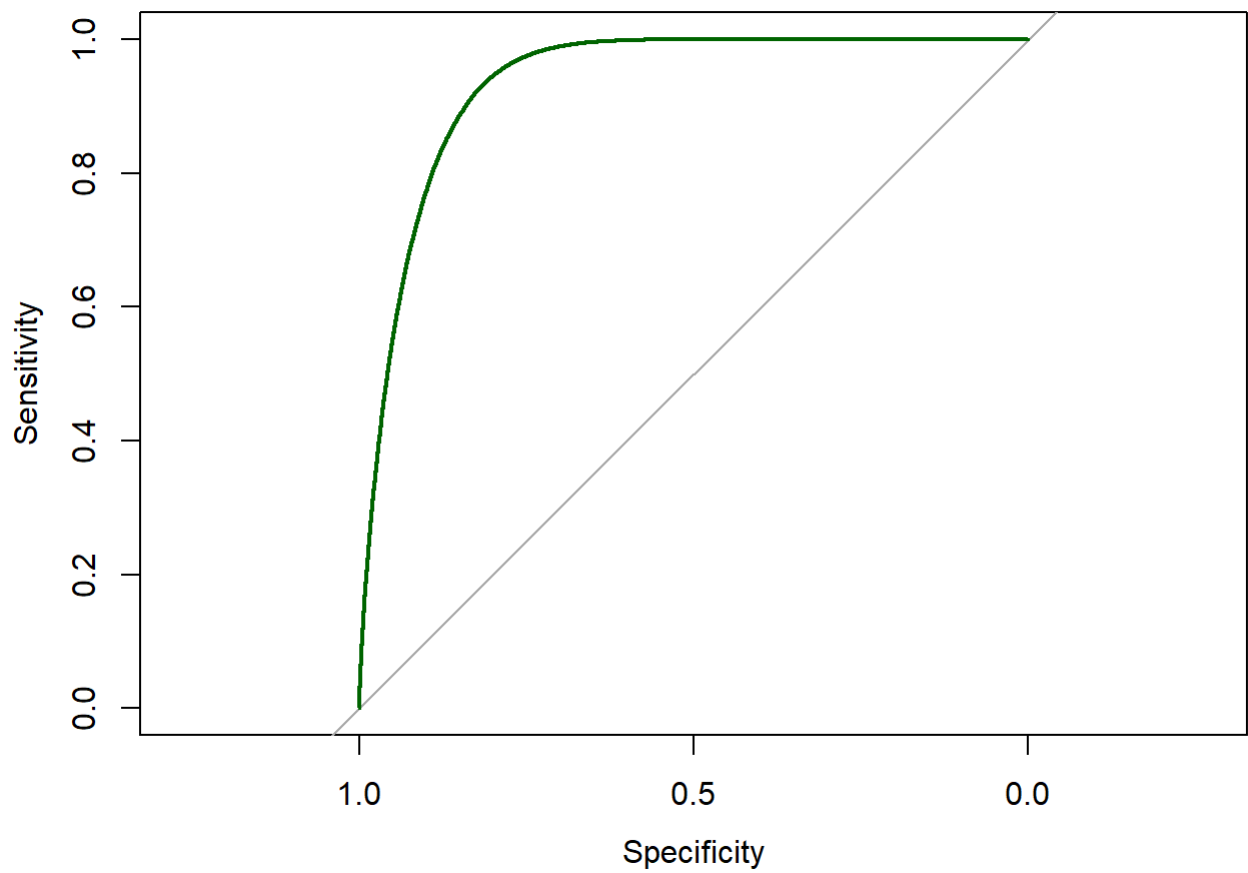
```
## Setting direction: controls < cases
```

Confusion Matrix and Statistics

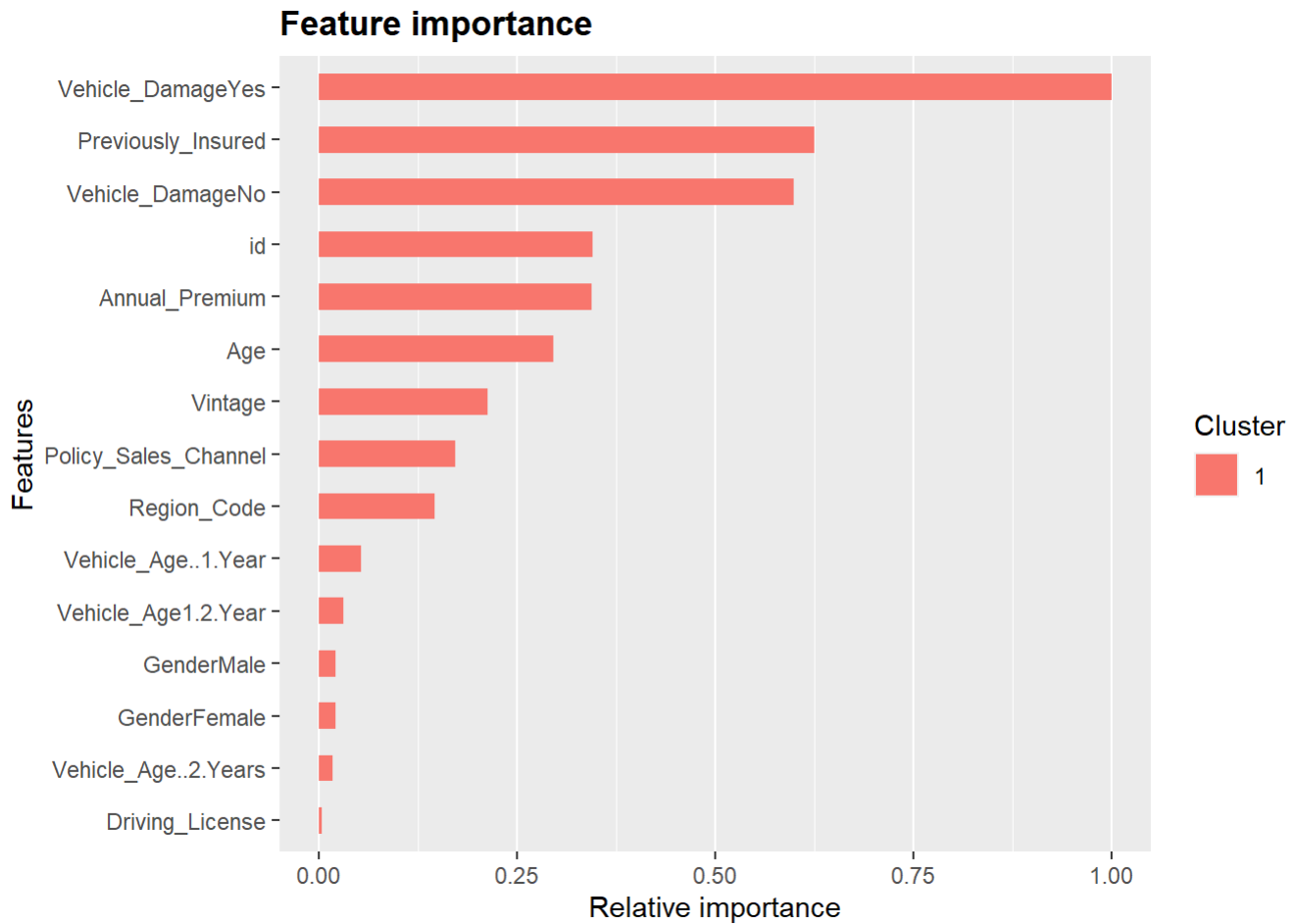


	Statistics
Accuracy	0.86
Kappa	0.71
AccuracyLower	0.86
AccuracyUpper	0.86
AccuracyNull	0.5
AccuracyPValue	0
McnemarPValue	0

ROC Curve



```
## $`Features Importance`
```



```
##
## $`Confusion Matrix`
## TableGrob (2 x 2) "arrange": 3 grobs
##   z    cells  name                grob
## 1 1 (2-2,1-1) arrange      gtable[layout]
## 2 2 (2-2,2-2) arrange      gtable[rowhead-fg]
## 3 3 (1-1,1-2) arrange      text[GRID.text.963]
##
## $`Roc Curve`
##
## Call:
## roc.default(response = label_train, predictor = cv$pred, levels = c(0, 1))
##
## Data: cv$pred in 190444 controls (label_train 0) < 190665 cases (label_train 1).
## Area under the curve: 0.9348
```

Submission (Checking Results)

Submit XGB predictions

	id <int>	Response <int>	predicted_response <dbl>	real_predictions <dbl>
1	381110	0	0	1.623483e-05
2	381111	0	0	3.630469e-01
3	381112	0	0	2.555647e-01
4	381113	0	0	2.934246e-03
5	381114	0	0	6.002499e-07
5 rows				

Model Deployment

Once we've selected a model, we proceed to deploy it with a data pipeline into a production or production-like environment for final user acceptance. This prepares the model for seamless integration with the client's existing applications. We successfully deployed our model and presented it to Triks Insurance Ltd.

From our initial understanding of the business, we identified several ways in which the client aims to benefit from this project:

- Providing GOOL Auto with insights into the potential business Triks Insurance can generate from their existing client base.
- Equipping Triks Insurance staff with a tool to prioritize clients for targeted marketing campaigns.
- Enabling Triks Insurance to project future revenues.
- Offering real-time notifications for potential cross-selling opportunities.

Pipelines can be developed for each of these requirements, although they fall outside the scope of this project. Predictions can be generated in real-time or on a batch basis. We have deployed the model to predict cross-sells based on user-provided customer input, including gender, age, region code, policy sales channel, vehicle age, and vehicle damage. Additionally, another deployment facilitates data file uploads, allowing predictions for entire datasets.

Conclusion

To achieve better accuracy we performed a hyper parameter tuning of the other parameters such (learning rate), max_depth (maximum depth of tree), sub sample values, etc by using a hyper grid and then achieved a better tuned mode.