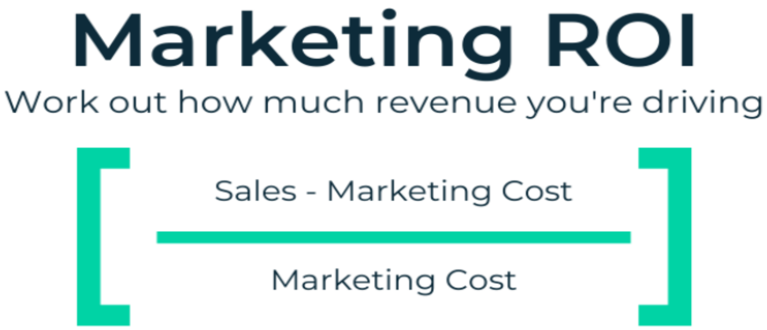
# Strategic Marketing Using Calibrated Probabilistic Classifiers for Insurance Firms

Daniyal Shahzad

## Introduction

In the landscape of the insurance industry, marketing departments have traditionally approached budget allocation with a broad stroke, distributing resources evenly across all customer segments. This conventional approach, while ensuring a level of outreach to a diverse audience, may not harness the full potential of strategic marketing. This points towards the potential of strategic marketing that can generate huge revenues for the company.

At the heart of this budget allocation problem lies the fact that not all customers are equal when it comes to the likelihood of purchasing insurance products. In an era of data-driven decision-making, there exists a tremendous opportunity to optimize marketing return on investment (ROI) by identifying and strategically targeting those customer segments with the highest likelihood of conversion. By doing so, insurance firms can move beyond the one-size-fits-all model of budget allocation and, instead, embrace a more personalized and effective engagement strategy.



This report explores the untapped potential within insurance marketing departments to enhance their ROI through the adoption of calibrated probabilistic classifiers. These classifiers not only prediction for individual customers, but also allow for the identification of customer segments that exhibit a higher propensity to purchase insurance products. By leveraging this insight, marketing teams can tailor their communication strategies, whether through personalized emails, targeted advertisements, or other forms of direct engagement, to resonate with the specific needs and preferences of these high-potential customer segments.

I aim to shed light on the impact of moving beyond uniform budget distribution and embracing a dynamic strategy that aligns with the unique characteristics of each customer segment. The potential benefits include not only increased conversion rates but also a more efficient use of marketing resources, ultimately fostering a more resilient and adaptive marketing approach in the competitive landscape of the insurance industry.

## Problem

We want to sell vehicle insurance to our current customers who already have life insurance. To do this well, we need to figure out the best way to spend our marketing budget. The aim is to get the most out of our budget by reaching the customers who are most likely to be interested in buying vehicle insurance. This can be done using personalized ads, or incentivized strategies. The challenge is to intelligently decide where to focus the marketing efforts to connect with existing policyholders who are most likely to be interested in this new insurance. With data and smart planning, it is possible to make the marketing efforts more efficient and successful.

## Motivation

The parallel nature of risk predictions and targeted marketing is evident in probabilistic modeling. A customer's likelihood of purchasing insurance is intrinsically linked to their risk profile, and thus, a unified approach to both aspects holds promise for achieving comprehensive insights. By extending the application of calibrated probabilistic classifiers beyond marketing optimization, I uncover a multifaceted solution that transcends individual functions within an insurance company.

The versatility of calibrated probabilistic classifiers extends beyond risk and marketing. The exploration of their application in predicting customer Lifetime Value (LTV) and utility in financial reporting highlights the potential of such models. By integrating calibrated probabilities into these functions, an insurance company can not only optimize customer targeting but also refine its financial strategies, ensuring a more adaptive organizational framework.

This report aims to bridge the gap between traditional marketing practices and cutting-edge data science methodologies, offering an integrated model that can serve as a catalyst for innovation within the insurance industry. By understanding the symbiotic relationship between marketing, risk, and financial predictions, I aspire to equip insurance companies with a tool that goes beyond targeted customer engagement, enhancing their overall resilience and performance in an increasingly dynamic market environment.

## Data Overview and Exploratory Analysis

Link: Insurance Dataset

For this project, I will utilize **a highly imbalanced dataset available on Kaggle**. This dataset represents customers of an insurance firm that offers health insurance. The dataset is structured to help answer the question of whether a policyholder would also be interested in purchasing vehicle insurance from the same company. The target variable consists of two classes, with the minority class representing less than 20% of the entire dataset. Furthermore, this data allows me to readily answer the question of how to predict the probability of a sale for each customer, which can then allow the insurance company to allocate resources towards those who are more likely. This data represents a real-life problem which is common yet not tackled optimally. Lastly, in real life, we come across datasets like these since the data is sensitive and is usually encrypted, therefore having little or no description.

The dataset consists of features that cover demographics, vehicle, and policy characteristics. The data contains 382,154 observations. According to the EDA, the datasets do not have any missing values.

In total, there are ten independent variables and one response variable. The details are given below.

| Feature | Description | Type | Levels |
|---|---|---|---|
| Sex | the sex of a customer | Categorical | 2 |
| Age | the age of a customer | Numerical | |
| Driving_license | Indicates whether the customer posses a driving license | Categorical | 2 |
| Region_code | The region a customer belongs to (Ranging from 0 to 52) | Categorical | 53 |
| Previously_Insured | Indicates whether a customer previously had a vehicle insurance | Categorical | 2 |
| Vehicle_Age | the age of a customer's vehicle | Categorical | 3 |
| Vehicle_Damage | Indicates if the customer's vehicle is damaged or not | Categorical | 2 |
| Annual_Premium | Amount of premium a customer will pay if they buy a vehicle insurance (Currency undefined) | Numerical | |
| Policy_Sales_Channel | The unique identifier of referral agency | Categorical | 156 |
| Vintage | The number of days the customer has been associated with the company | Numerical | |
| Response | Indicates whether customer purchased a vehicle insurance | Categorical | 2 |

The absence of numerical features remains a potential problem with the data as it can affect the models' predictive performance. Furthermore, having most of the features as categorical can introduce the curse of dimensionality due to one-hot encoding. This shows that LASSO regularization would be a suitable approach to handle large number of redundant categories. Additionally, many of the categories are rare, which means applying LASSO regularization would lead to the model simply dropping these categories resulting in a change of interpretation for that specific feature. Lastly, this also adds to visualization challenges since numerical features are more easily studied using scatterplots and correlation matrices.

Given the sensitive nature of the insurance industry and the proprietary information held by individual firms, details regarding the specific origin of the dataset are intentionally withheld to maintain the confidentiality and privacy of the data sources.

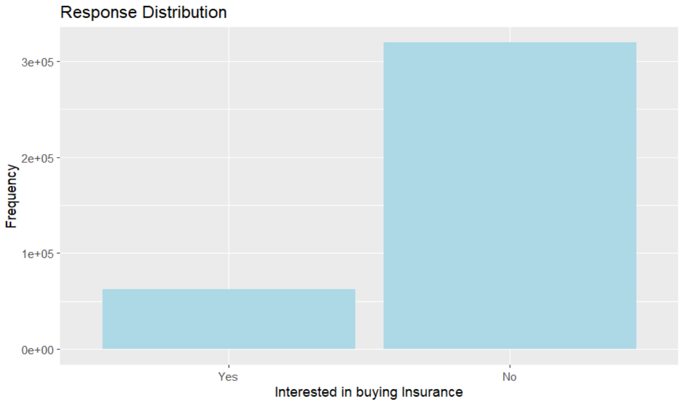
It is important to note that while the specific source remains undisclosed, the characteristics and composition of this dataset are representative of a typical dataset found within the broader insurance industry. The intentional abstraction of identifiable details ensures that any insights derived from the analysis are applicable to a wide range of insurance firms. This emphasis on generalizability enhances the external validity of the findings, allowing for broader applicability and relevance to the industry at large.

This dataset poses unique computational challenges, primarily stemming from the prevalence of categorical variables, with some featuring an extensive range of categories, surpassing 150. The abundance of categorical data introduces complexity in terms of computational efficiency and demands specialized techniques for analysis.

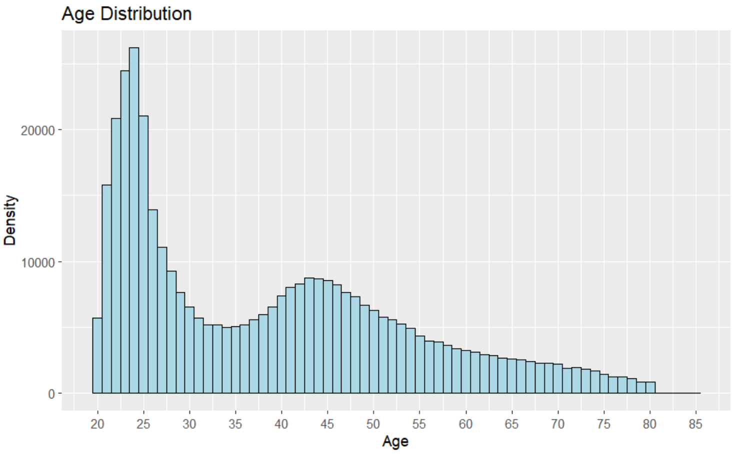
## Exploratory Data Analysis

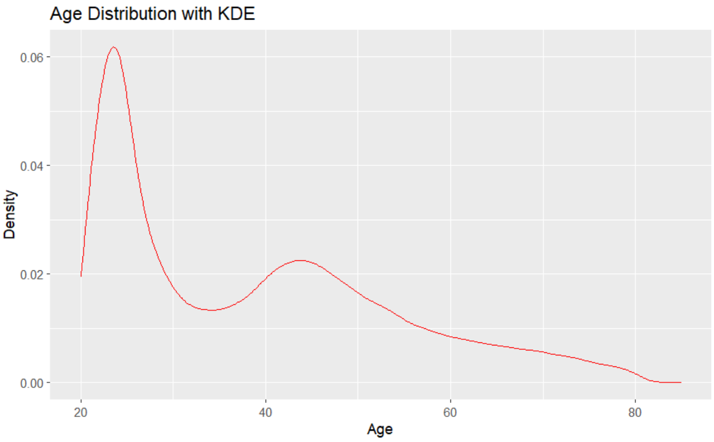
### Univariate Analysis

The **response** variable, as highlighted before, is highly imbalanced with less than 20% of the customers being interested in buying the insurance.

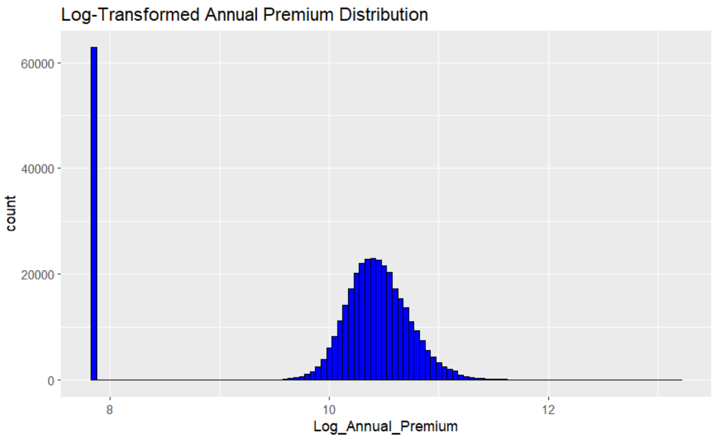


Eyeballing the **Age** distribution, I can see that most of the customers are of age 25 and around. It tells me that most of the customers are young. There are also a significant number of customers in the range of 40-50. Using the IQR method and 1.5 * IQR as the distance, I find there are outliers in Age, customers with age of 85. Age is an interesting variable because younger people are more likely to be in a car accident, but also since they are young and less wealthy, they are less likely to buy insurance. Maybe this variable can be transformed using a log transformation since the distribution looks right skewed.


Age Distribution


Age Distribution with KDE

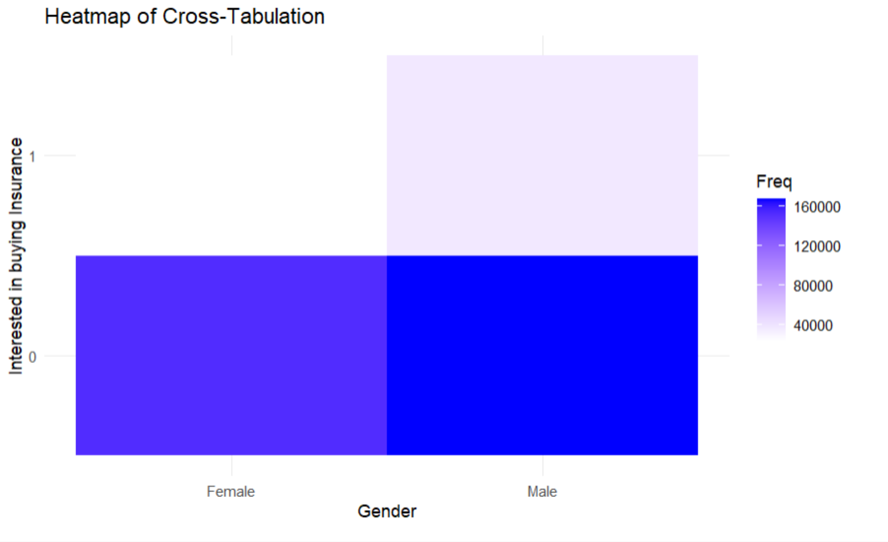
The **Annual Premium** represents the amount the customer will pay for the vehicle insurance per year. Plotting the histogram for annual premium showed very weird results so I resorted to plotting Kernel Density Estimates, and a violin plot. All of them showed the data was highly skewed (Right). The reason to transform is that the original feature has extreme values which might influence the predictive power of the model and if I select variable, this variable might get chosen because of the extreme effect from its extreme values. It's important to note that now the interpretation of the coefficient of this new feature has changed. Note that the bar at 7.9 represents the default values given to a customer if their annual premium was missing, a common practice in the industry.


Log-Transformed Annual Premium Distribution

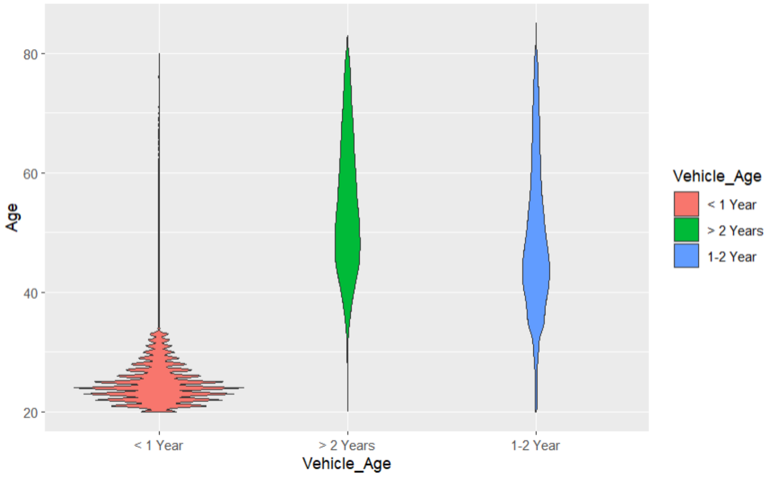
## Bivariate Analysis

Due to the nature of the dataset and there being a majority of categorical, it is a challenge to visualize all the features. Yet using domain knowledge, I highlight a few important ones. Furthermore, since independence of features is not an assumption for logistic regression, I have not tested categorical features' independence.

Plotting a heatmap of **Sex Against Response** uncovers various insights. At first sight, it seems males are way more likely to be interested in buying vehicle insurance. But since Sex shouldn't really be correlated with interest in insurance, that means there are likely confounding variables present that influence this observed sex-response relationship which requires further investigation.



Most of the newer cars are owned by younger people. Which contradicts the general notion of young people not being rich enough to afford new cars. This again can point towards confounding variables. Also shows some existence of multicollinearity.



## Feature Importance

Random forests are somewhat robust to class imbalance due to bootstrapping. Therefore, to top it off, I have used Random Forest to explore feature importance. Vehicle Damage, and Previously Insured comes out to be the most important whereas Sex, and driving license seems to be the least, something that was already expected after the analysis.

## Literature Review

To optimize the Return on Investment (ROI) from the marketing strategy, it is crucial to recognize that the data will likely be highly imbalanced. This imbalance arises because reaching a considerable number of individuals is necessary to convert only a small proportion of them into actual sales. Frank E. Harrell, a prominent figure in statistics, elaborates in his blog on how logistic regression can effectively handle class imbalance when the model is correctly specified.

While the previous work on this specific problem seems to be sparse but a plethora of papers can be found on binary classification under class imbalance. The paper “Portfolio creation using artificial neural networks and classification probabilities: a Canadian study” comes close to our problem of using classification probabilities to optimize allocation of marketing budget.

Dealing with a high-class imbalance in the training data, one of the most popular approach has been SMOTE discussed in the paper "SMOTE: Synthetic Minority Over-sampling Technique". The paper has 2,426 citations but has been found to lead to classifier mis-calibration, leading to suboptimal forecasts and resource allocation, pointing towards the gap between academia and actual applications of models. An insightful paper, "The Harm of Class Imbalance Corrections for Risk Prediction Models: Illustration and Simulation Using Logistic Regression," delves into the potential mis-calibration issues arising from imbalance corrections which is very relevant. Lastly, "GHOST: Adjusting the Decision Threshold to Handle Imbalanced Data in Machine Learning" which offers insights into threshold adjustment techniques which even though is not detrimental to the calibration, but results in loss of information since two customers who are likely to buy vehicle insurance will both be tagged as 1, where as in the case where one of them was more likely, this information would be lost.

Calibration of classifiers is important in the vehicle insurance industry, particularly under imbalanced dataset where the rare event generates huge revenues for the company. The paper "Approaches for Credit Scorecard Calibration: An Empirical Analysis" showcases various methods for calibrating probabilistic classifiers since not all classifiers output calibrated probabilities by default. Papers such as "On Calibration of Modern Neural Networks" shed light on the lack of calibration in modern neural networks and the need for calibration techniques. In this project, it is vital for the insurance company to determine expected Marketing spending ROI to enable fair and effective resource allocation.

The paper “Unsolved Problems in ML Safety” explains how calibrated probabilitisitc predictions can ensure safety in autonomous vehicles, highlighting the importance of forecasts that can easily be verified and are based on real life events.

For evaluation of such probabilistic models, it is impertinent to use proper scoring rules since these rules will provide valid metrics to judge the performance. The paper “A note on the use of empirical AUC for evaluating probabilistic forecasts” highlights the use of popular metric ROC AUC is incorrect but is used widely. For a proper scoring rule, I use Brier score as suggested by the paper “Verification of Forecasts Expressed in Terms of Probability”.

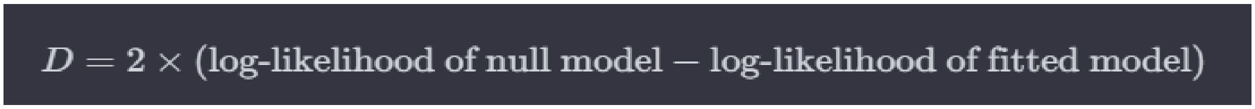
# Methodology

The data undergoes an initial split into training and testing sets, employing a stratified approach to ensure consistent class ratios in both datasets. This split adheres to an 80-20 distribution, with the test data serving as "unseen" data during the subsequent model evaluation. Moving on to preprocessing, dummy variables are generated for all categorical variables. It's noteworthy that while popular R libraries can automatically create dummy variables, their capability is limited by the maximum number of levels a categorical variable can handle, with the maximum for random forest set at 53. To maintain consistency, dummy variables are manually created for all categorical variables.

Informed by insights from the Exploratory Data Analysis (EDA), a log transformation is applied to the Annual Premium variable, effectively addressing any outliers within this feature.

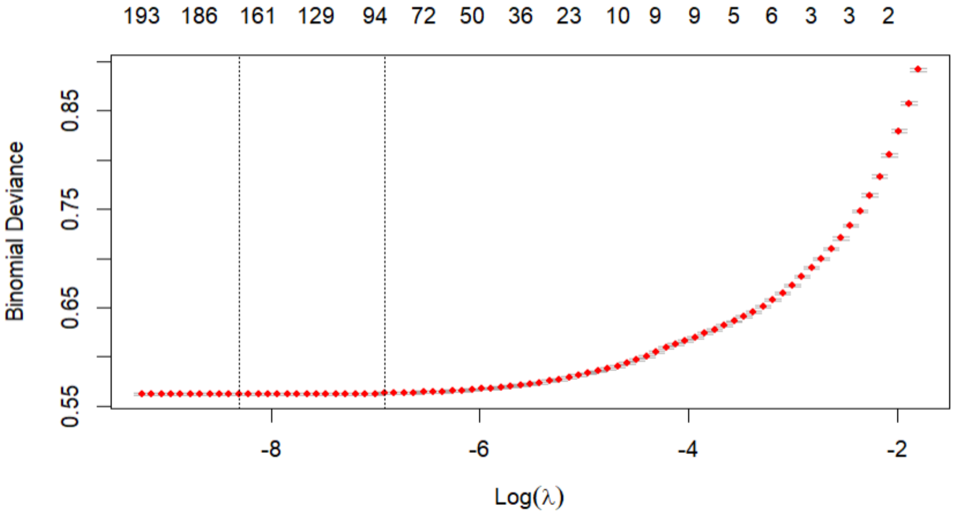
Following the principles discussed in class, a binary logistic model is fitted to the data using logistic regression. Given the presence of a substantial number of categories, LASSO regularization is incorporated to eliminate redundant categories, preventing overfitting and ensuring a feasible training process, considering the large number of categories and observations.

For the selection of the optimal lambda, cross-validation is employed, and a lambda value is chosen to minimize the Deviance, a methodology extensively covered in our course. The deviance is defined as follows:



$$D = 2 \times (\text{log-likelihood of null model} - \text{log-likelihood of fitted model})$$

The graph for Deviance against Log(Lambda) highlights the optimal lambda chosen for our model which is then used in training the final model.



Lastly, we went through bootstrapping extensively in class, and using the same idea prediction intervals were created, by resampling from the data with replacement until we have the same size. Then training the model and predicting for each test observations. These prediction intervals provide the confidence of our model for each individual predictions.

## Evaluation of the Model

Receiver Operating Characteristic Area Under the Curve (ROC-AUC) is a widely used metric for evaluating the performance of binary classification models. While ROC-AUC is informative in many scenarios, it is considered an improper scoring rule in the context of class imbalance.

ROC-AUC measures the ability of a model to rank instances in terms of their predicted probabilities. It is insensitive to the absolute values of the predicted probabilities and only considers their rank order. This property makes ROC-AUC less sensitive to changes in class probabilities. Furthermore, it can be insensitive to class imbalance, especially when one class is rare. It evaluates the model's ability to discriminate between positive and negative instances but does not directly penalize misclassification of the minority class. In highly imbalanced datasets, a model can achieve a high ROC-AUC by simply assigning high probabilities to the majority class, even if it performs poorly on the minority class.

A proper scoring rule should be optimized when the predicted probabilities are close to the true probabilities. ROC-AUC, being focused on rank ordering, does not directly measure the calibration of predicted probabilities. In a proper scoring rule, a model is incentivized to provide accurate probability estimates, which is important for decision-making, especially in applications with imbalanced classes.

It is worthy to note that scoring rules like F1-score can be used here but since we are interested in probability forecasts and not classification, this will not be used.

The Brier Score, also known as the mean squared error, is a proper scoring rule that evaluates the accuracy of probabilistic predictions. It is often recommended as a more appropriate metric than ROC-AUC in our scenario where accurate probabilistic predictions are necessary for budget allocation. A model that outputs well-calibrated probabilities will receive a lower Brier Score.

Unlike ROC-AUC, the Brier Score is sensitive to the absolute values of predicted probabilities. It penalizes models for being overconfident or underconfident, making it more informative about the calibration of probability estimates.

Lastly, it has a straightforward interpretation as a mean squared error. A lower Brier Score indicates better predictive accuracy, and the score can be easily understood in terms of the model's average deviation from the true outcomes.

The Brier Score is given by,

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

Where,

N is the number of observations, ft is the probability forecast for a given customer and Ot is the binary response

Something noteworthy here is that Brier Score is indeed affected by a class imbalance. For example, if our probabilistic model forecasts 0 for every customer (the majority class), then our brier score comes out to be 0.164 on the testing set. This will become my baseline score.

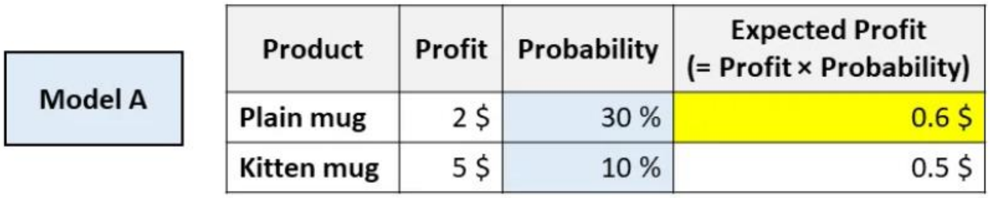
Lastly, Brier Skill Score can be used to see the % improvement over a baseline model. This is given by,

$$BSS = 1 - \frac{\text{Brier Score of the Forecast Model}}{\text{Brier Score of the Reference Model}}$$

Calibration

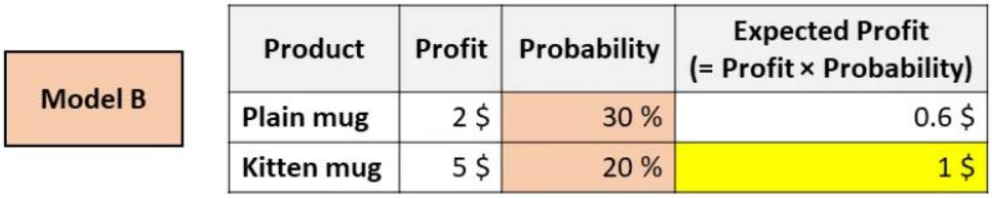
Calibration is an important aspect of probabilistic predictions, especially in applications where the predicted probabilities represent confidence levels or probabilities of events occurring. A well-calibrated model provides probability estimates that align closely with the true frequencies of events, enhancing the interpretability and reliability of the predictions. Calibration is crucial in scenarios where decision-making is based on predicted probabilities, such as risk assessment, medical diagnosis, or credit scoring.

An example to illustrate the importance of calibration is given below,



Model A

| Product | Profit | Probability | Expected Profit (= Profit × Probability) |
|---|---|---|---|
| Plain mug | 2 $ | 30 % | 0.6 $ |
| Kitten mug | 5 $ | 10 % | 0.5 $ |

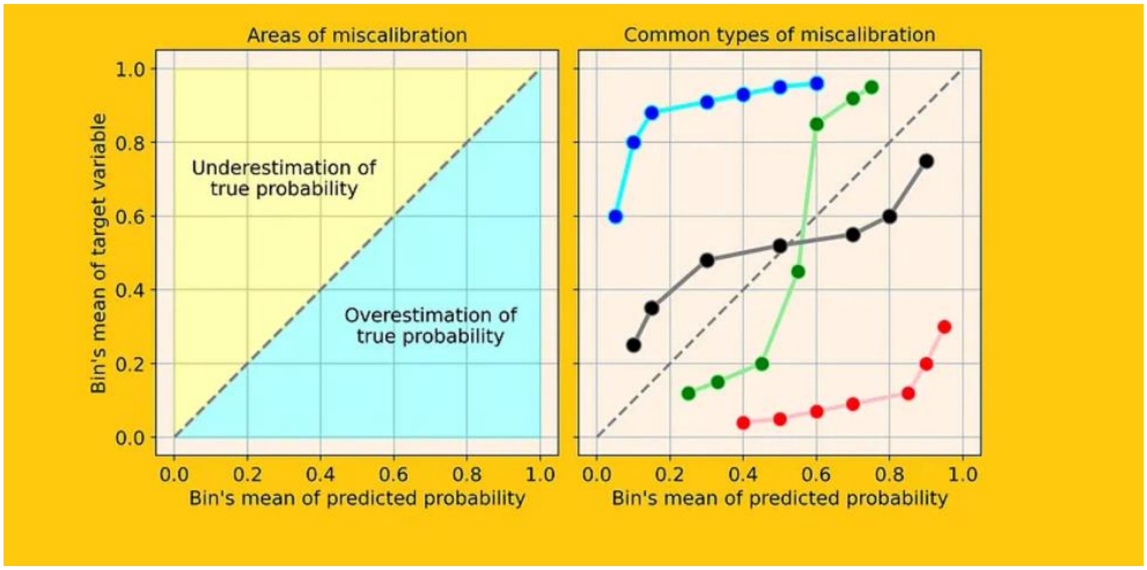


Model B

| Product | Profit | Probability | Expected Profit (= Profit × Probability) |
|---|---|---|---|
| Plain mug | 2 $ | 30 % | 0.6 $ |
| Kitten mug | 5 $ | 20 % | 1 $ |

From the figure above, it can be seen that even though both the model have the same ROC-AUC, both these models output different result. Model A predicts that the Plain Mug would be more profitable while Model B predicts that the Kitten Mug is more profitable.

To come to our final decision, it is imperative that we compare these probabilities with the ground truth, the data in our case. This is where calibration curves come in, as they compare the predicted probabilities with the actual proportion of events observed in a set of observations.



Graphical Representation: Calibration curves plot the predicted probabilities against the observed frequencies of events. A well-calibrated model will have points close to the diagonal line (y = x). Deviations from the diagonal indicate areas where the model may be underestimating or overestimating probabilities.

Calibration curves are sensitive to miscalibrations, making it easy to identify areas where a model may need improvement. This visual inspection is especially valuable for identifying systematic biases in predictions.

## Justification

The rationale behind selecting logistic regression stems from its ability to yield well-calibrated probabilities, especially when contrasted with other models. Additionally, its computational efficiency is noteworthy, a crucial factor given the substantial size of our dataset, where lengthy training periods would be impractical. The model not only offers interpretability but, with the incorporation of LASSO regularization, efficiently manages redundant categories and guards against overfitting. Collectively, these considerations make logistic regression the preferred modeling technique.

## Findings and Results:

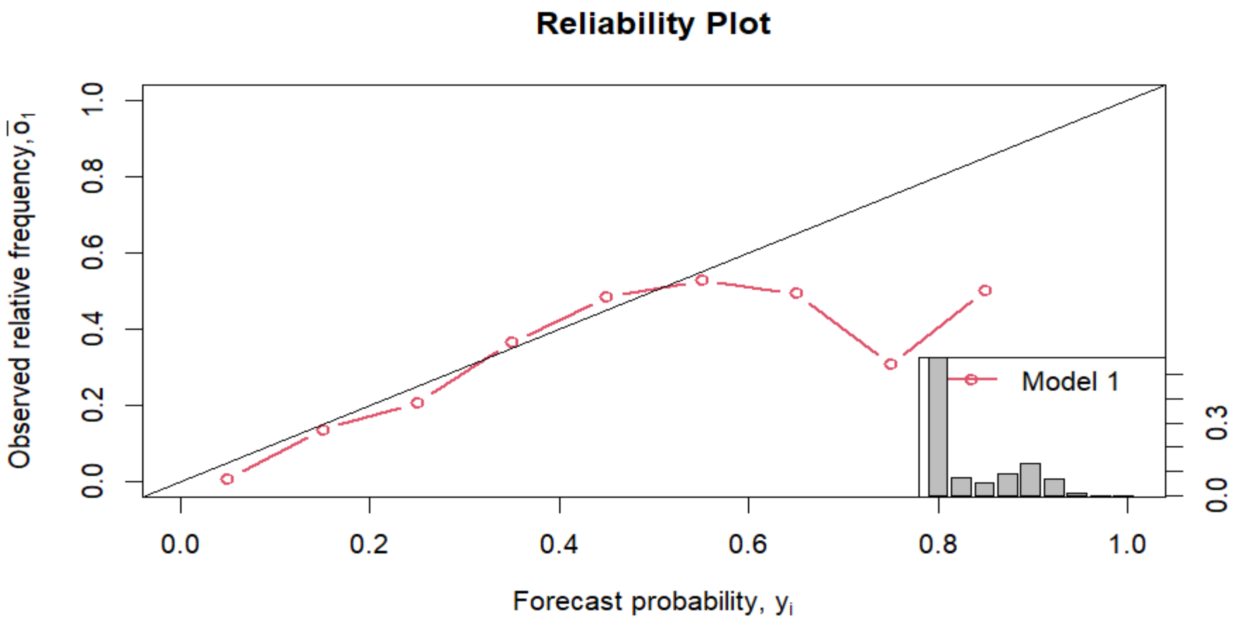
The evaluation of our Logistic Regression shows promising results. Its brier scores are given in the table below.

| Logistic Regression | Training Data | Testing Data |
|---|---|---|
| Baseline BS | 0.163821 | 0.1637708 |
| Logistic BS | 0.09537485 | 0.09531107 |
| Brier Skill Score | 0.418 | 0.418 |

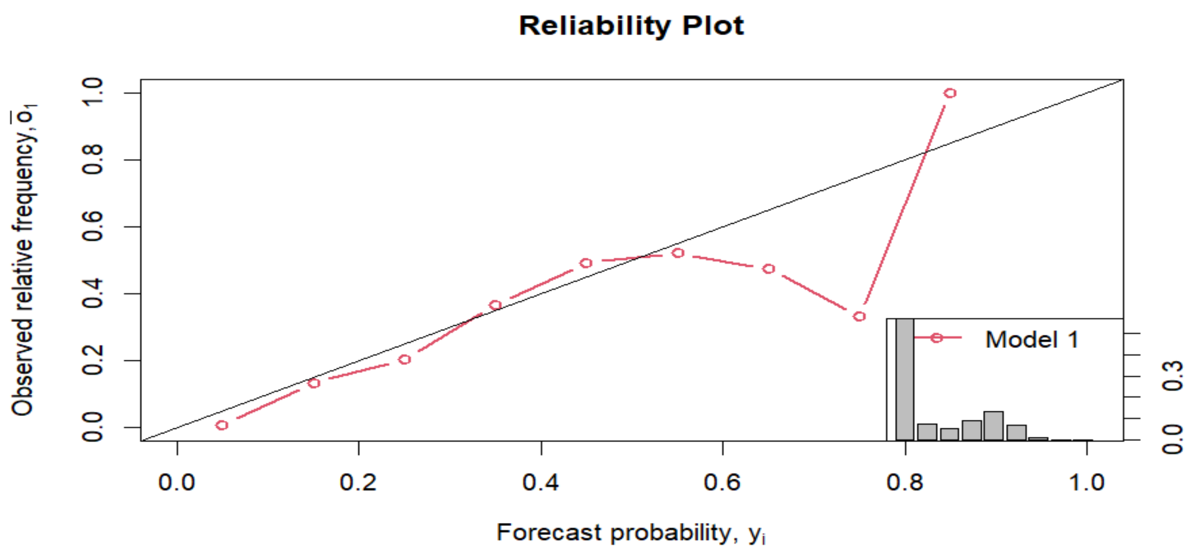
The model is performing about 42% better than the baseline model. Secondly, the brier scores seem to be very similar across training data and testing data, indicating that the model is generalizing well.

Moving on to the calibration curves,

Calibration Curve for the training data seems to be well calibrated. It is important to note that Last three bins seems to be uncalibrated, this is due to very small sample in those bins making their estimates very volatile. Since majority of the customers are in bin 1 to bin 7, we can conclude that it is sufficiently calibrated on the training data.



For the testing data, the calibration curve follows the same trend, with last 3 bins having very low sample (<1% of the total observations), resulting in a very high variance. Additionally, it seems to be sufficiently calibrated for the bins 1 to 7 since they contain the majority of the customers.

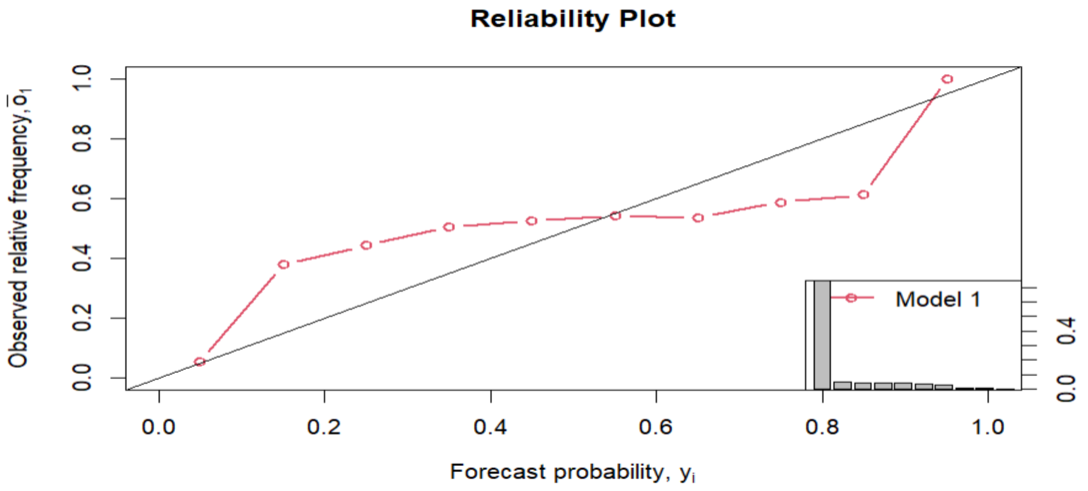


It is important to note that the reason why initial bins have most of the observation is because of the the class imbalance in the data.

To further study the adequacy of the model, a Random Forest was also trained the same way and its results recorded.

| | Training Data | Testing Data |
|---|---|---|
| Logisitic BS | 0.09537485 | 0.09531107 |
| RF BS | 0.1123852 | 0.1085347 |
| Brier Skill Score Against Logistic Regression | -0.1771 | -0.1385 |

This underscores that random forest performs worse than our logisitic regression, by about 14%. The finding contributes to the adequacy of our selected model. A result that was expected since Random Forest calculates probability scores based on the proportion of trees that predict a sale. This is not a probability, but instead a score and hence is mis-calibrated.



The lower performance can be seen using its calibration curve, which provides a sigmoidal shape of calibration curve. A property of Random Forests. It is evident that random forests is under predicting for bins 2 to 4, and over predicts afterwards. Also, the last three bins have very low number of observations, hence the miscalibration there can be ignored because of high variability. This finding favors our chosen model, that is logisitic regression.

Lastly, Neural Networks were also intended for this analysis but due to the computational limitations given the size of the data, ANNs have not been explored for our problem.

## Findings

Our objective is to identify customers or customer segments strategically targetable to maximize return on investment. The model successfully discerned the characteristics of customers most likely to opt for vehicle insurance, yielding several key insights:

1) Customers with undamaged vehicles exhibit a higher likelihood of opting for vehicle insurance.
2) Individuals without prior vehicle insurance are also more inclined to consider purchasing.
3) There is an increased chance of sales among customers whose life insurance was procured through specific sales channels, such as phone calls or booked appointments.
4) Residence in particular regions correlates with a higher likelihood of purchasing vehicle insurance, possibly indicating affordability due to residing in affluent areas.
5) Ownership of newer vehicles significantly enhances the likelihood of vehicle insurance sales.
6) As anticipated from the Exploratory Data Analysis (EDA), gender plays a minimal role in determining the likelihood of a sale.
7) Vintage, akin to gender, has a negligible coefficient, indicating its limited relevance.

Moreover, given that the model predicts the probability of a customer purchasing vehicle insurance, these probabilities can be employed as weights for the allocation of the marketing budget among customers. Specifically, customers with a high probability of buying vehicle insurance would receive a higher allocation, facilitating the implementation of more personalized marketing strategies. In essence, the model adeptly identifies customers who are prime targets for strategic marketing efforts, thereby implicitly ensuring the optimal utilization of the marketing budget.

### Utility of the Model

Once we have a model that can predict the probability of a sale, it can be utilized not only for the current problem but for many others.

For our current problem, the insurance company was initially allocating all its marketing budget equally across customers. This can include emails, TV ads, flyers etc. without targeting any customer segments. This model informs the company which segment is more likely to buy vehicle insurance so that they can make personalized marketing and appeal to customers on either a segment level or an individual level. For example, making calls to customers likely to buy vehicle insurance or incentivized strategies to boost sales among those customers who are less likely to buy vehicle insurance can boost Return on Invest of the marketing budget.

The utility of this model extends to supplemental financial reporting as well. Since insurance firms are required to recognize their future profit revenues in their financial reports, to get this future revenue, our model can be utilized to generate expected sales forecast because the output of the model is probabilities and not probability scores, we can use this to get expected Revenue by multiplying the output probability by vehicle insurance total policy premium.

Finally, by utilizing the constructed prediction intervals, we can identify customer segments characterized by 'low variability' or 'high variability.' For instance, customers classified as 'low variability' may not need significant marketing efforts, as they likely already have a predetermined inclination. Conversely, 'high variability' customers represent a potential opportunity for successful conversion and merit focused marketing strategies. This was not under the scope of this report but presents an interesting problem.

## Conclusion

In summary, this study has successfully identified key attributes influencing the likelihood of vehicle insurance purchases, offering actionable insights for strategic marketing decisions. By leveraging a predictive model, we highlighted the significance of factors such as vehicle condition, prior insurance history, effective sales channels, geographical location, and vehicle age. The model's capacity to discriminate between customer segments enables a shift from uniform marketing budget allocation to targeted, personalized strategies, maximizing the probability of conversion and optimizing return on investment of the marketing budget.