Vehicle Insurance Predictor

Shaili Gandhi
AU1841012
Ahmedabad University
Ahmedabad, India
shaili.g@ahduni.edu.in

Priyank Sangani
AU1841136
Ahmedabad University
Ahmedabad, India
priyank.h@ahduni.edu.in

Yash Arvindkumar Patel
AU1841141
Ahmedabad University
Ahmedabad, India
yash.p4@ahduni.edu.in

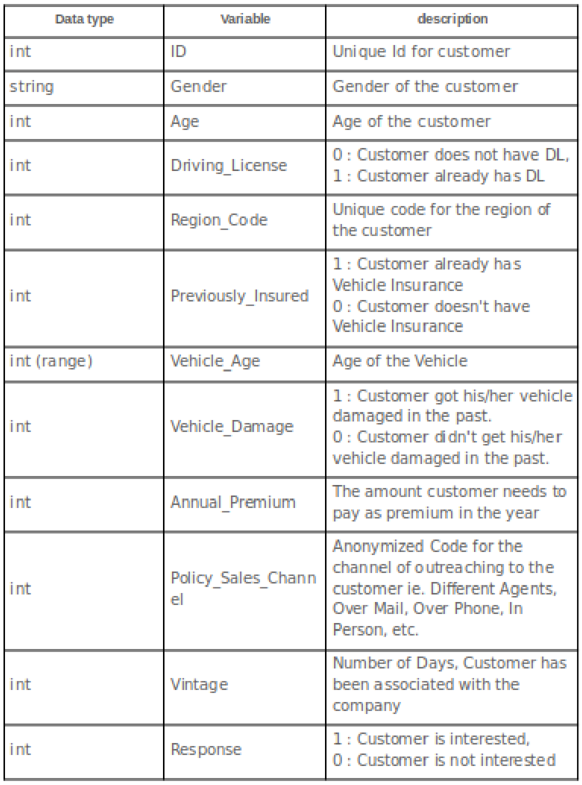
Samarth Shah
AU1841145
Ahmedabad University
Ahmedabad, India
samarth.s@ahduni.edu.in

***Abstract*—In the second part of the project, there is model creation using Scikit learn library for logistic regression classifier and Support Vector Classifier as well as model from scratch for logistic regression and performance matrices based on prediction of all the models.**

***Index Terms*—Logistic Regression, Support Vector Machine, Scratch, Scikit learn, Performance**

I. Introduction

An insurance policy is an arrangement in which the company agrees to provide a compensation in case of specified damage, in return for payment of a specified premium amount. Since, a lot of people pay the premium, but only a few of them face vehicle accidents and damage, and get the compensation, everyone shares the risk of everyone else. So, the insurance business is based on the existence of risks and the desire to avoid them. Thus, the data based quantification of risk and uncertainty plays a crucial role in this field, and this is how machine learning comes into the picture.



Data type	Variable	description
int	ID	Unique Id for customer
string	Gender	Gender of the customer
int	Age	Age of the customer
int	Driving_License	0 : Customer does not have DL, 1 : Customer already has DL
int	Region_Code	Unique code for the region of the customer
int	Previously_Insured	1 : Customer already has Vehicle Insurance 0 : Customer doesn't have Vehicle Insurance
int (range)	Vehicle_Age	Age of the Vehicle
int	Vehicle_Damage	1 : Customer got his/her vehicle damaged in the past. 0 : Customer didn't get his/her vehicle damaged in the past.
int	Annual_Premium	The amount customer needs to pay as premium in the year
int	Policy_Sales_Channel	Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.
int	Vintage	Number of Days, Customer has been associated with the company
int	Response	1 : Customer is interested, 0 : Customer is not interested

So, we are aiming to build a model that would predict whether a customer would be interested in buying the insurance based on the information like:

- Demographics: Gender, Age, Vehicle code
- Vehicle: Vehicle Age, Damage
- Policy: Premium

This would help the insurance company in optimizing its business model and revenue.

II. Literature Survey

The changing nature of data and the means of handling it is making the insurance industry more data driven and customer driven and is affecting their strategies. In analytical insurance problems, machine learning tools are specially designed to address issues like pricing, claims management, marketing and prevention. This article focuses on developing a dynamic pricing system for online motor vehicle liability insurance. The data for this prediction was obtained by collecting and monitoring competitors' products for a long period of time. the response variable is the premium offered for a given customer profile. The features used where car characteristics, age of the driver, date of driving permit, etc. Many different machine learning algorithms were used for the modelling like gradient boosting models, deep learning models, random forests, etc. cross validation was used to optimise the values of the hyper parameters for each model. The insurers regularly change and adapt their tariffs, so the predictions need to be monitored on a regular basis. Thus, without machine learning dynamic pricing would not have been possible. Ultimately, substantial gains can be obtained from the high model production quality and the speed of implementation.[11] Major insurance companies are focusing on excavating useful knowledge and information hidden in users, products and services in the massive customer data and acquiring more customer resources and thus gain new competitive advantage. This article focuses on the feature selection using methods like random forest, gradient lifting tree and lifting machine algorithm. after analysing the data it can be known if the features have little influence on whether or not to renew the insurance. here the classification accuracy rate is used as a model classification performance evaluation index. The results show that the light GBM algorithm has a better classification effect.[12]

III. Implementation with Mathematical Analysis

We took help of normalization for preprocessing There has been some mathematical equations which have used for data cleaning i.e. for data Normalization. After that we need to design a mathematical intuition for classification algorithms. Till now we understood the mathematical intuition for logistic regression.

A. Analysis for Data cleaning and visualization

The main motive for data normalization is to convert the numeric values of the features into common range i.e 0 to 1, without changing differences in the ranges of values.
There are four columns Region Code, Annual Premium, Policy Sales Channel and Vintage.

- For Vintage and Regional Code, it is observed that only 2 & 3 values are mostly used in the dataset. While the other values are less used and that is why data normalization is chosen.
$$Normal = \frac{Actual - minimum}{maximum - minimum}$$
- For Annual Premium, it is inferred that data is highly skewed towards left side.
- For Policy Sales Channel, the graph is uniformly distributed and that is why data normalization is chosen for that column.

B. Analysis for Logistic Regression

In Logistic Regression, sigmoid function is used as hypothesis and loss is calculated equation mentioned below. Here Gradient Descent is used to calculate the optimal theta parameters.

$$hypothesis = \frac{1}{1 + e^{-(\theta^T X)}}$$

$$Loss = -\frac{1}{m}((y)log(h) + (1-y)log(1-h))$$

Here for Loss Function, if the actual label is 0, then the second term in the equation remains and and if the label is 1 then only the first term in the equation will remain and accordingly, 1 is subtracted from the hypothesis.
Then Gradient Descent is carried out to minimize loss.

$$\theta^{new} = \theta^{old} - \alpha * \frac{\partial loss}{\partial \theta}$$

Here alpha is the learning rate which defines the rate at which our model will learn. For the scratch part initially, the alpha is taken as 0.001.

C. Analysis for Performance Matrix

The Confusion matrix is used to evaluate the classification model. The Confusion matrix is a 2x2 matrix.
where C_{ij} is equal to the number of observations known to be in group i and predicted to be in group j.
Thus in binary classification, the count of true negatives is C_{00}, true positives is C_{11}, false negatives is C_{10} and false positives is C_{01}.

Precision is a performance metric used to answer how many positive values are predicted correctly among all the correctly predicted values.

$$Formula: \frac{TP}{TP + FP}$$

Here TP: True Positive & FP: False Positive

Recall is a performance metric used to answer how many actual positive values are predicted among all the actually positive values.

$$Formula: \frac{TP}{TP + FN}$$

Here TP: True Positive & FN: False Negative

Accuracy is a performance metric used to answer how many correctly predicted values are there among all the predicted values.

$$Formula: \frac{TP + FP}{TP + TN + FP + FN}$$

Here TP: True Positive, TN: True Negative, FP: False Positive & FN: False Negative

The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. ROC Curve is a probability curve which plots the True Positive Rate(TPR) vs False Positive Rate(FPR) at various threshold values.

IV. Experiments and Results

In this section we will discuss the experiments we have done and respective results and try to draw inference from them.
Experiments are based on performance of different model and on performance of different hyper-parameters.

A. Comparison between Classifiers

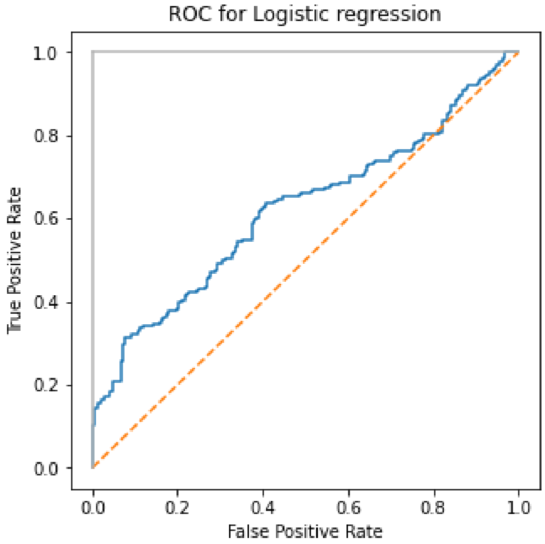
To predict whether the customer will be interested in buying vehicle insurance or not, we need to use classifier to train the data. We tried three models for this:

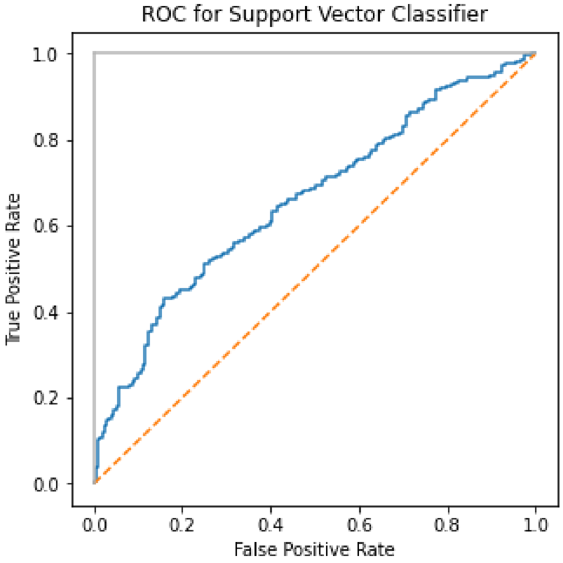
- Logistic Regression using Scikit Learn
- Logistic Regression from scratch
- Support Vector Machine using Scikit Learn

Here is the table of performance of different models:

Models	Accuracy	Precision	Recall
Logistic Regression (sklearn)	0.61	0.68	0.64
Logistic Regression (Scratch)	0.41	-	-
SVM (sklearn)	0.62	0.68	0.66

After getting the values of confusion matrix we also plotted ROC curve from True Positive Rate and False Positive Rate





Inferences from the above experiment:

- Here we can see that models with sklearn seems to have larger accuracy then scratch model.
- For both the sklearn models, we can say that accuracy is almost similar, also precision value is similar as well and for recall value both of them have a very minor difference.
- For ROC, we can determine whether the ROC curve is good or bad by looking the Area under the curve and we can see that ROC for Logistic Regression is converging with the diagonal line after 0.8 point where as ROC for SVM has convergence when it reaches 1.
- So clearly the Area under the Logistic Curve is lesser than the area under the SVM curve.

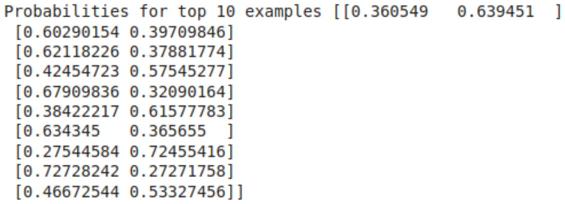
Now we will go through the final step which we have carried out is the confidence probabilities for both classes while testing the models. We have drawn these probabilities for Logistic regression and SVM models.
Confidence probabilities are for each example in test dataset so there will be large number of them but here we are showing the images of some of them.
When we get the trained weights and bias, we need to pass it through the sigmoid function to get the class and after getting the output from sigmoid function we will compare it with 0.5 i.e the threshold value. When the values will be greater than 0.5 then it will be considered as class 1 and if less than 0.5 then it will be considered as class 0.

For Confidence probabilities we have used

$$P(x/y) = \frac{1}{(1 + exp^{W*f(x)+B})}$$

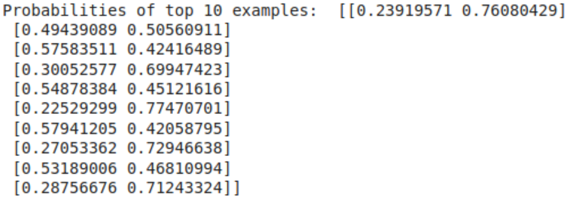
where, f(x) is a function of data points to calculate the distance from the hyperplane and W are weights and B is bias.

Confidence Probabilities for Logistic regression



```
Probabilities for top 10 examples [[0.360549   0.639451  ]
 [0.60290154 0.39709846]
 [0.62118226 0.37881774]
 [0.42454723 0.57545277]
 [0.67909836 0.32090164]
 [0.38422217 0.61577783]
 [0.634345   0.365655  ]
 [0.27544584 0.72455416]
 [0.72728242 0.27271758]
 [0.46672544 0.53327456]]
```

Confidence Probabilities for SVM



```
Probabilities of top 10 examples:  [[0.23919571 0.76080429]
 [0.49439089 0.50560911]
 [0.57583511 0.42416489]
 [0.30052577 0.69947423]
 [0.54878384 0.45121616]
 [0.22529299 0.77470701]
 [0.57941205 0.42058795]
 [0.27053362 0.72946638]
 [0.53189006 0.46810994]
 [0.28756676 0.71243324]]
```

B. Comparison based on Hyperparameters

After experimenting with classifiers, we have experimented with the hyperparameters for logistic regression using sklearn and SVM using sklearn models.
We have experimented with penalty term that is whether it is l1 or l2. Penalty l1 is Let's first start with Logistic Regression :

penalty	C value	accuracy	precision	recall
l2	0.5	0.61	0.69	0.64
l1	0.1	0.60	0.68	0.64
l2	0.5	0.60	0.68	0.64
l1	0.1	0.61	0.69	0.62

- Here we can see that all the models have almost same accuracy just the difference occurs in the precision column.
- Also here we have shown 2 decimals, there are differences in 3rd decimal and hence the differences are very small.

Below is the Results of Accuracy using Support Vector Machine with different kernels and with different C values.

Kernel	C value	accuracy
Linear	0.1	0.61
Linear	0.5	0.75
Linear	1	0.75
Polynomial	0.1	0.75
RBF	0.1	0.75

- Here we have experimented with the kernels and we can see that no major differences are reported except linear kernel and 0.1 c value.
- For the rest models, the differences are too small that they are not seen when we consider only two decimals.

V. Conclusion

- At first we have gone through the introduction where we introduces our data and all the features.
- Then we listed the literature review where we gone though some existing work.
- Then we gone through detailed mathematical analysis for our models and performance matrix.
- After comparing Logistic Regression and SVM Model for our dataset. It is observed that SVM Model performs better than Logistic Regression in terms of accuracy.
- As we know that there has been margin which ensures the robustness of the model provides higher accuracy for SVM model.
- Then we experimented with the hyperparameters i.e. kernel, value of C, Penalty term.
- It is observed that in most of the cases, hyperparameter changes doesn't lead to more accuracy.
- For our dataset, SVM with polynomial kernel and Radial Basis Function(RBF) Kernel with C value equal to 0.1 yields similar accuracy.

References

[1] Wang, Hui Dong. "Research on the Features of Car Insurance Data Based on Machine Learning." Procedia Computer Science 166 (2020): 582-587.

[2] Grize, Yves-Laurent, Wolfram Fischer, and Christian Lützelschwab. "Machine learning applications in nonlife insurance." Applied Stochastic Models in Business and Industry 36.4 (2020): 523-537.

[3] Mane, Sandeep and Srivastava, Jaideep and Hwang, San-Yin and Vayghan, Jamshid,2004 12,475- 478,Estimation of false negatives in classification, 0-7695-2142-8, 10.1109/ICDM.2004.10048.

[4] https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

[5] https://towardsdatascience.com/understanding-logistic-regression-step-by-step-704a78be7e0a

[6] https://www.altexsoft.com/blog/datascience/machine-learning-project-structure-stages-roles-and-tools/

[7] https://towardsdatascience.com/data-visualization-for-machine-learning-and-data-science-a45178970be7

[8] https://towardsdatascience.com/how-to-tackle-any-classification-problem-end-to-end-choose-the-right-classification-ml-algorithm-4d0becc6a295

[9] https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

[10] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[11] https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.2543

[12] https://www.sciencedirect.com/science/article/pii/S1877050920301381