Project Report

Regression Analysis - Caravan Insurance Policy Purchase Prediction

Regression and Time series Analysis Class Fall 2023

Name	NetID
Layakishore Reddy Desireddy	LD786
Nancy Soni	NS1583
Chetan Reddy Valluru	CV410
Vikram Sriram Yasaswi Sagaram	VS865

Abstract

The objective of this machine learning project is to develop and implement a predictive model that can accurately forecast whether which customer characteristics are important to predict whether a customer is likely to purchase the insurance policy.

By analysing historical customer data and identifying key patterns and correlations, the model aims to provide reliable predictions about customer behaviour. This will enable more efficient targeting and personalization of insurance products, ultimately enhancing customer engagement and increasing sales effectiveness.

The project seeks to leverage advanced machine learning techniques to gain deeper insights into customer preferences and decision-making processes, facilitating more informed and strategic business decisions in the insurance domain.

The project utilises hypothesis testing and other statistical analysis to reduce computational and human resources expenditure while retaining the rate of prediction for target customers.

Introduction

In this project, our objective was to construct a robust predictive model to identify potential customers for the Caravan Policy. We analyzed a spectrum of policyholder information, including Customer Type, age, number of houses, relationship status, and existing policies. To achieve this, we employed a diverse set of regression approaches such as Logistic Regression, Deep Learning, Neural Network, Random Forest Regression, and Forward Stepwise Regression. Our emphasis was on accuracy, experimenting to discover the most fitting model for our data. This brief provides a snapshot of our exploration into predictive modeling, offering valuable insights for enhancing Caravan Policy customer acquisition strategies.

Data (https://archive.ics.uci.edu/static/public/125/insurance+company+benchmark+coil+2000.zip)

Dataset Description

We utilized the dataset from UCI Machine Learning Repository, specifically the "Insurance Company Benchmark (COIL 2000)" dataset. The data contains 9823 real customer records. Each record consists of 86 variables, containing sociodemographic data (variables 1-43) and product ownership (variables 44-86). The sociodemographic data is derived from zip codes. All customers living in areas with the same zip code have the same sociodemographic attributes. Variable 86 (Caravan) indicates whether the customer purchased a caravan insurance policy. Among total 9823 customer data, 586 customers have purchased the policy and 9236 customers have not purchased the policy.

Data Overview

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	ORIGIN	MOSTYPE	MAANTHUI	MGEMOMV	MGEMLEEF	MOSHOOFD	MGODRK	MGODPR	MGODOV	MGODGE	MRELGE	MRELSA	MRELOV	MFALLEEN	MFGEKIND	MFWEKIND	MOPLHOOG	MOPLMIDD	MOPLLAAG	MBERHOOG	MB
2	train	33	1	3	2	8	0	5	1	3	7	0	2	1	2	6	1	2	7	1	
3	train	37	1	2	2	8	1	4	1	4	6	2	2	0	4	5	0	5	4	0	
4	train	37	1	2	2	8	0	4	2	4	3	2	4	4	4	2	0	5	4	0	
5	train	9	1	3	3	3	2	3	2	4	5	2	2	2	3	4	3	4	2	4	
6	train	40	1	4	2	10	1	4	1	4	7	1	2	2	4	4	5	4	0	0	
7	train	23	1	2	1	5	0	5	0	5	0	6	3	3	5	2	0	5	4	2	
8	train	39	2	3	2	9	2	2	0	5	7	2	0	0	3	6	0	4	5	0	
9	train	33	1	2	3	8	0	7	0	2	7	2	0	0	5	4	0	3	6	2	
10	train	33	1	2	4	8	0	1	3	6	6	0	3	3	3	3	0	1	8	1	
11	train	11	2	3	3	3	3	5	0	2	7	0	2	2	2	6	0	4	5	2	
12	train	10	1	4	3	3	1	4	1	4	7	1	2	0	3	6	4	3	3	0	
5824	test	33	1	4	2	8	0	6	0	3	5	0	4	1	1	8	2	2	6	0	
5825	test	6	1	3	2	2	0	5	0	4	5	2	2	1	4	5	5	4	0	5	
5826	test	39	1	3	3	9	1	4	2	3	5	2	3	2	3	6	2	4	4	2	
5827	test	9	1	2	3	3	2	3	2	4	5	4	1	2	4	4	2	4	4	2	
5828	test	31	1	2	4	7	0	2	0	7	9	0	0	0	6	3	0	0	9	0	
5829	test	30	1	2	4	7	1	4	2	3	5	0	4	4	3	2	1	2	6	1	
5830	test	35	1	2	4	8	2	5	1	2	8	0	1	2	5	3	1	5	4	2	
5831	test	6	1	3	3	2	3	4	2	2	9	0	0	0	5	4	4	4	2	4	
5832	test	4	1	2	4	1	0	7	2	0	9	0	0	1	7	2	3	4	2	2	
5833	test	10	1	4	2	3	0	7	0	2	9	0	0	0	2	7	2	3	5	0	
5834	test	34	1	3	3	8	0	9	0	0	5	0	4	2	3	4	0	3	6	0	
5835	test	36	1	3	3	8	2	1	0	6	8	0	1	1	1	7	1	5	4	2	
5836	test	38	1	3	2	9	0	6	3	0	9	0	0	0	4	5	2	4	4	2	
5837	test	11	1	2	3	3	1	3	1	5	4	2	4	4	2	3	1	3	6	1	

Data Dictionary

The data dictionary provided offers a detailed overview of a dataset used for predicting caravan insurance policy uptake. This dataset is likely used in the context of data science and analytics, particularly in the insurance industry. Below is an overview of the data, categorized into various domains:

1. Customer Demographics and Lifestyle (Variables 1-35)

Customer Subtype (MOSTYPE): Classifies customers into various groups based on income, family composition, and lifestyle (e.g., "High Income, expensive child", "Elderly singles", etc.). These are detailed in L0.

Housing and Household Composition (e.g., MAANTHUI, MGEMOMV): Information on the number of houses owned, average household size, and types of families (e.g., singles, families with/without children).

Age and Education (MGEMLEEF, MOPLHOOG, MOPLMIDD, MOPLLAAG): Average age groups, and levels of education ranging from high to low.

Social and Economic Status (MBERHOOG, MBERMIDD, MSKA, MSKB1, etc.): Classifies customers based on occupational status and social classes, indicating their economic positioning.

Housing Type and Vehicle Ownership (MHHUUR, MHKOOP, MAUT1, MAUT2, etc.): Information about whether the customer rents or owns a house, and the number of cars they own.

2. Insurance Contributions and Ownership (Variables 44-85)

Contributions to Various Insurance Policies (PWAPART, PBESAUT, PBRAND, etc.): Data on customer contributions to different types of insurance policies like car, life, fire, and disability insurances.

Number of Policies Owned (AWAPART, APERSAUT, ABRAND, etc.): Details the number of various insurance policies each customer holds, such as car, life, property, and social security insurances.

3. Insurance Policy for Caravans (Variable 86)

Caravan Insurance Policy (CARAVAN): Indicates whether the customer has purchased a caravan insurance policy, with values 0 (no policy) or 1 (policy purchased).

4. Categorical Value Labels (L0, L1, L2, L3, L4)

L0 - Customer Subtypes: Detailed classifications of customers based on sociodemographic factors like age, income, family status, and lifestyle.

L1 - Age Groups: Categorizes customers into age brackets ranging from 20 to 80 years.

L2 - Customer Main Type: Classifies customers into groups based on broader lifestyle and status categories like "Successful hedonists", "Average Family", "Conservative families", etc.

L3 - Religious Affiliation: Percentage range of customers' affiliation with the Roman Catholic religion.

L4 - Contribution Levels to Private Third-Party Insurance: Specifies the range of contributions to private third party insurance, from none to very high contributions.

Overview and Implications for Insurance Modeling

Sociodemographic Data: Key for understanding customer profiles, which can be pivotal in predicting the likelihood of purchasing caravan insurance.

Insurance Contributions and Ownership: Insights into customers' existing insurance portfolios can help in identifying those more likely to invest in additional policies like caravan insurance.

Targeting Strategy: Data enables the creation of targeted marketing strategies and personalized insurance products.

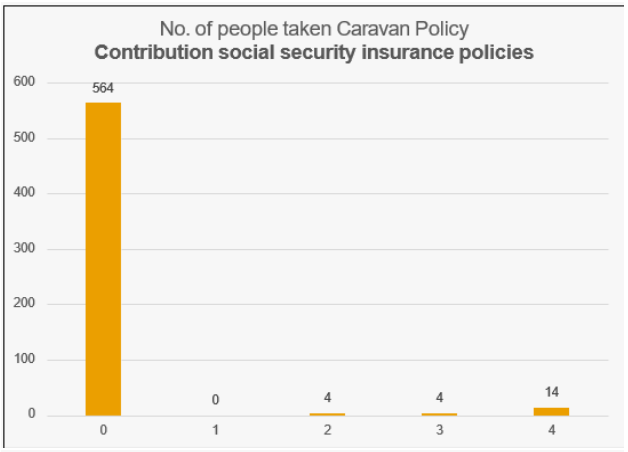
Regulatory Compliance and Portfolio Balance: Understanding customer profiles aids in maintaining regulatory compliance and a balanced insurance portfolio, avoiding concentration in high-risk groups.

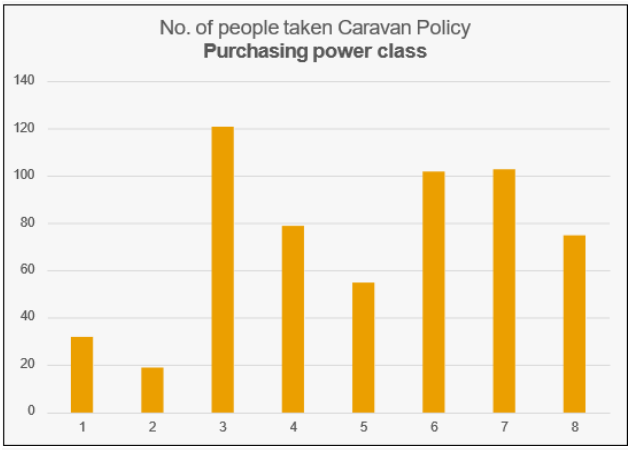
In summary, this dataset provides a comprehensive view of customers' sociodemographic profiles, insurance behaviors, and their likelihood to purchase caravan insurance, crucial for predictive modeling in the insurance sector.

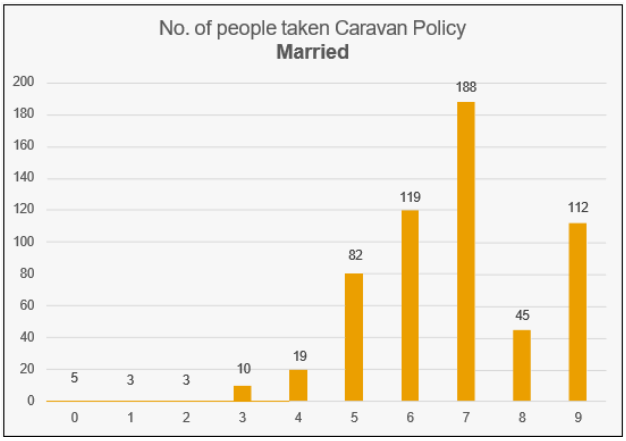
Data Preprocessing

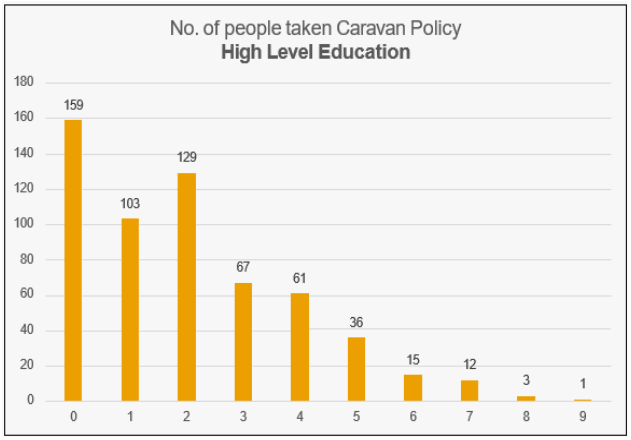
We tried various data split of the data and then split the data into a first 5822 customers as training set and rest 4000 as test set. We utilize it in the "ORIGIN" field to determine which row of data is for which set. There are then 85 other columns utilized as "features" of the dataset and the last (86^{th}) column is the target variable.

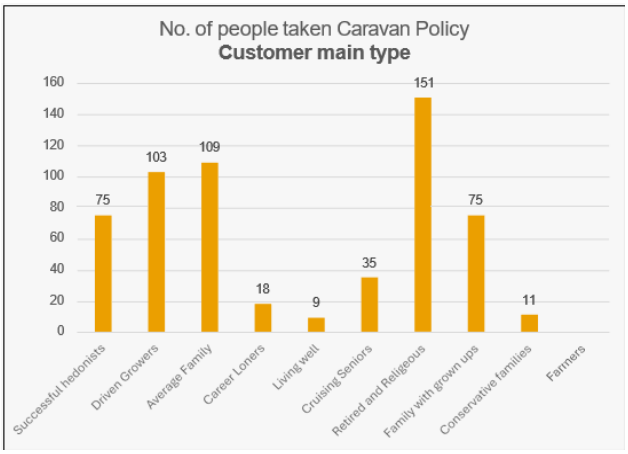
Exploratory Data Analysis

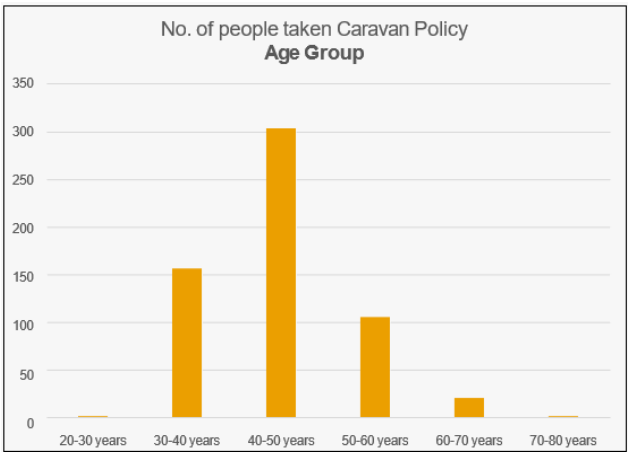




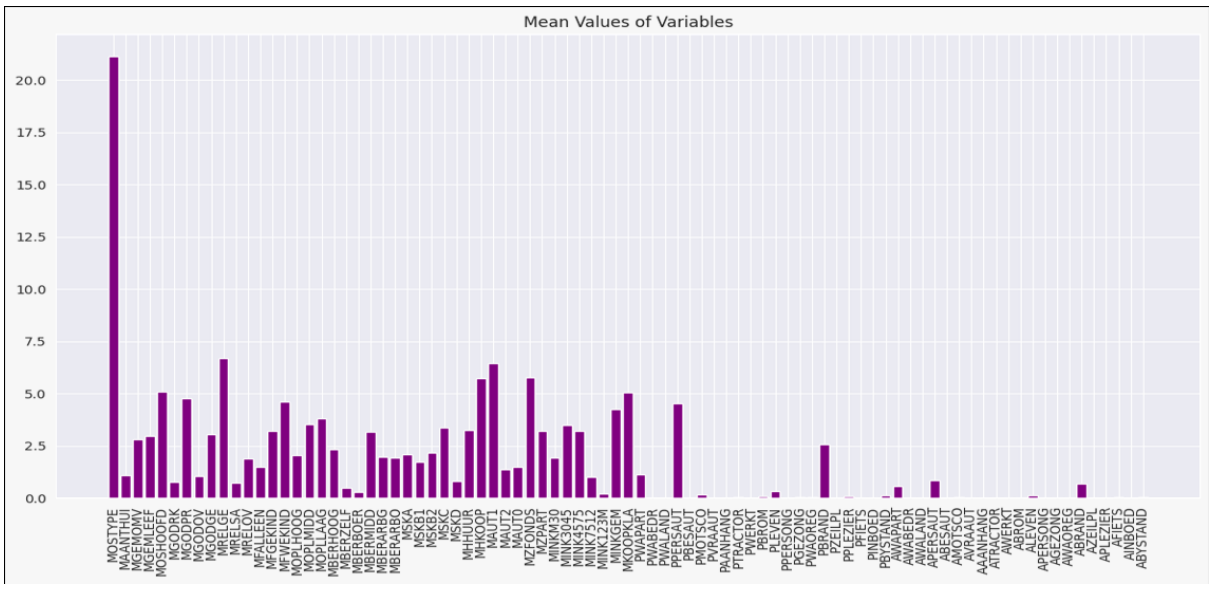








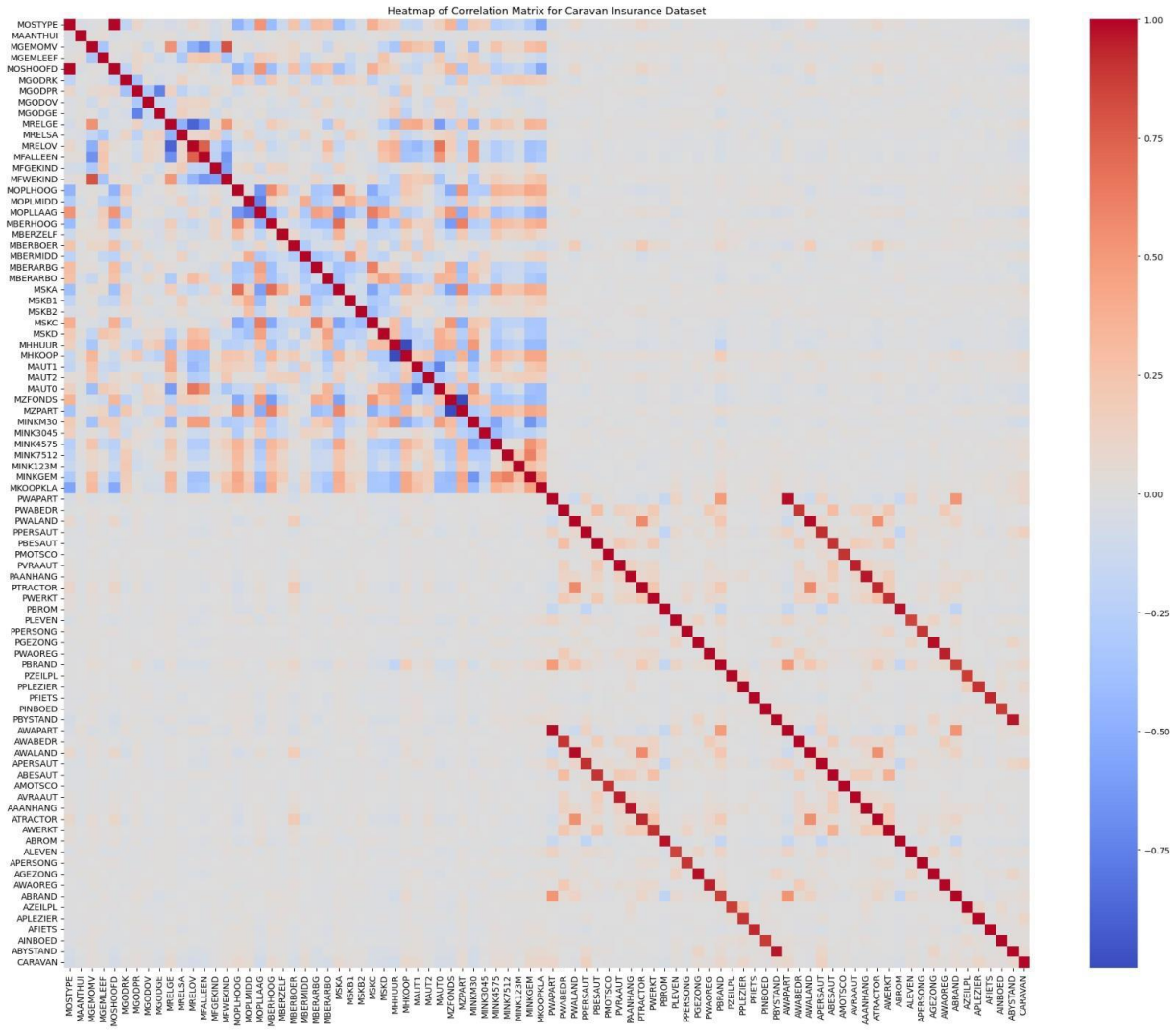
Mean values of variables who purchased Caravan Policy



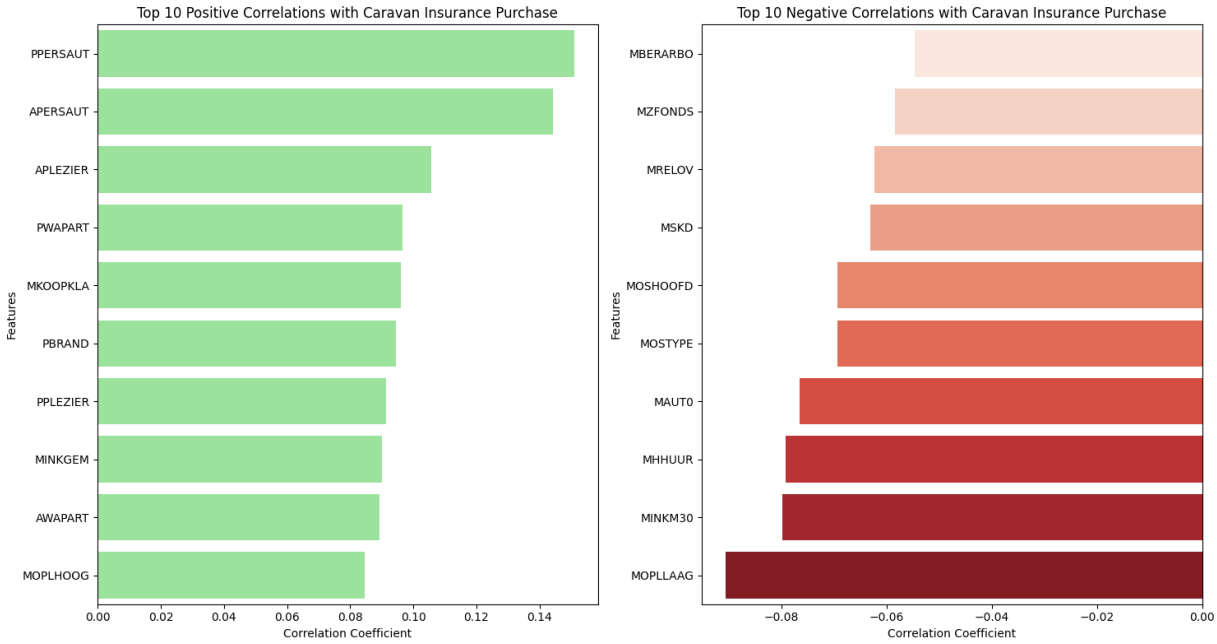
METHODOLOGY

Correlation Analysis

Correlation plot



From the correlation metrics we can see there is lot of multicollinearities I the data which is because the data is having two types of information 1. about whether the person owns the policy (for ex. Car policy) and 2. his contribution to that particular policy. Hence, we can deduce that there are redundant variables in our data. These redundant variables were seen removed after fitting different models, which shows the effectiveness of the fitted model.



Here are the top 20 negative and positive correlated features who have strongest linear relationship with our target variable. This graph tells us that person who owns car policy is most likely and person with low income and low education are most less likely to purchase the Caravan policy.

Modelling

- ✓ Logistic Regression with Full Model
- ✓ Neural Network with Full Model
- ✓ Stepwise Regression – Forward Selection
- ✓ Logistic and Neural Network with selected 16 features by Forward selection
- ✓ Random Forest Model to verify the effectiveness of above models

Model 1 : Logistic Regression with full model

What is Logistic Regression?

Logistic regression is a statistical technique used for binary classification. It predicts the probability of an event's occurrence by mapping data to a logistic (S- shaped) curve. This model relates a binary dependent variable to one or more independent variables, producing probabilities between 0 and 1. It's widely used in fields like healthcare, marketing, and machine learning for its efficiency in dichotomous outcome predictions.

data

	ORIGIN	MOSTYPE	MAANTHUI	MGEMOMV	MGEMLEEF	MOSHOOFD	MGODRK	MGODPR	MGODOV	MGODGE	...	APERSONG	AGEZONG	AWAOREG	ABRAND	AZEILPL	APLEZIER	AFIETS	AINBOED	ABYSTAND	CARAVAN
0	train	33	1	3	2	8	0	5	1	3	...	0	0	0	1	0	0	0	0	0	0
1	train	37	1	2	2	8	1	4	1	4	...	0	0	0	1	0	0	0	0	0	0
2	train	37	1	2	2	8	0	4	2	4	...	0	0	0	1	0	0	0	0	0	0
3	train	9	1	3	3	3	2	3	2	4	...	0	0	0	1	0	0	0	0	0	0
4	train	40	1	4	2	10	1	4	1	4	...	0	0	0	1	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9817	test	33	1	2	4	8	0	7	2	0	...	0	0	0	1	0	0	0	0	0	0
9818	test	24	1	2	3	5	1	5	1	3	...	0	0	0	1	0	0	0	0	0	1
9819	test	36	1	2	3	8	1	5	1	3	...	0	0	0	1	0	0	0	1	0	0
9820	test	33	1	3	3	8	1	4	2	3	...	0	0	0	0	0	0	0	0	0	0
9821	test	8	1	2	3	2	4	3	0	3	...	0	0	0	1	0	0	0	0	0	0

9822 rows × 87 columns

Evaluation of the Logistic Regression Model

Model Implementation

- Model Training: Logistic Regression model trained on standardized training data.
- Prediction: Model used to predict caravan insurance policy uptake on test data.
  - Training Data: Extracted from rows labelled 'train' in the ORIGIN column. We considered the first 5822 rows as training data.
  - Test Data: Extracted from rows labelled 'test' in the ORIGIN column. The remaining data is considered as test data.

- Features and Labels:
  - Independent Variables (Features): All columns except 'CARAVAN'.
  - Dependent Variable (Label): 'CARAVAN' column indicating caravan insurance policy uptake.

- The logit $f(x_1, x_2) = b_0 + b_1x_1 + b_2x_2 + \ldots\ldots$ (Linear Regression)
- Sigmoid Function: $p(x_1, x_2) = 1 / (1 + \exp(-f(x_1, x_2)))$
- Loss Function : Log loss or cross entropy

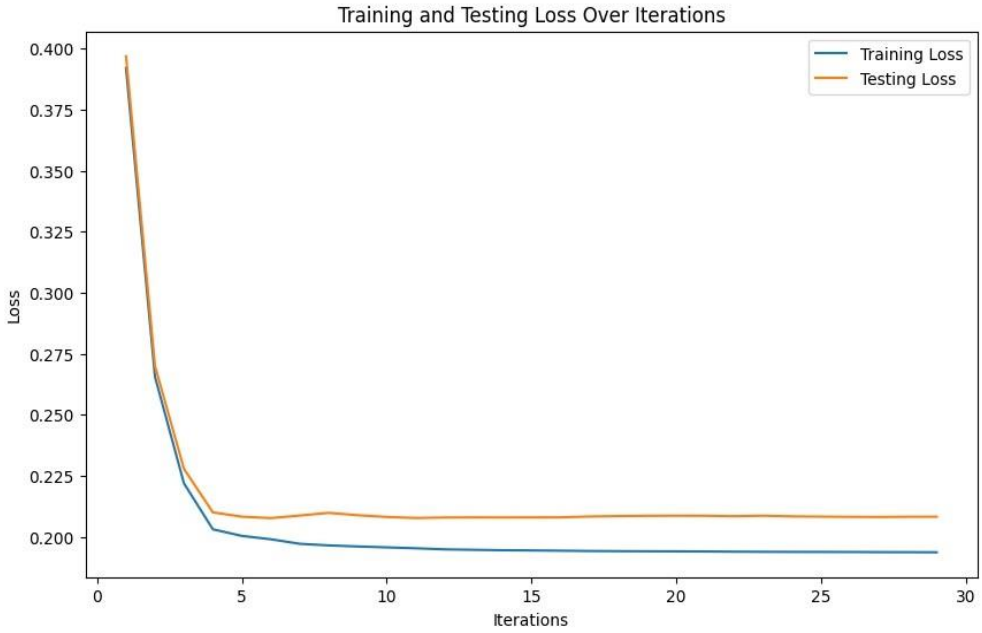
$$L(y,p) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)]$$

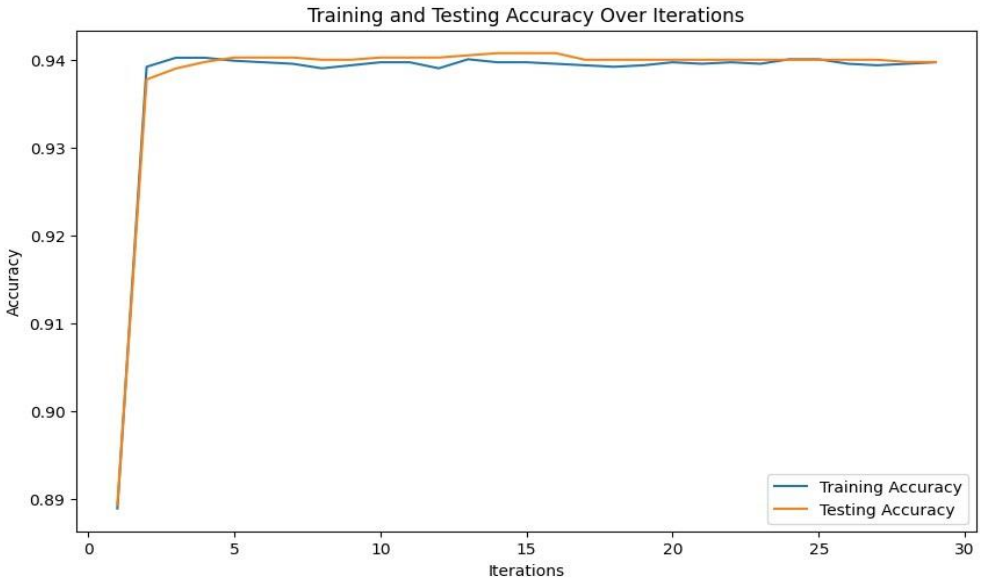
N is the number of observations.

yi is the actual label for observation i, which can be either 0 or 1.

pi is the predicted probability that observation *i* belongs to the class with label 1.

Gradient descent is used for back propagation.





Performance Metrics

Accuracy Score:

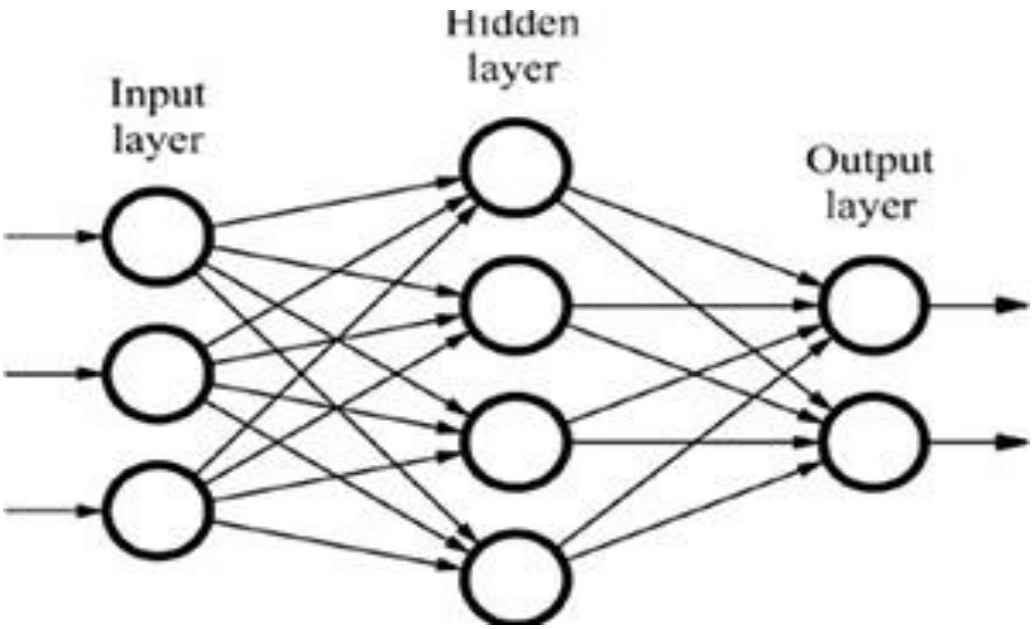
- Value: 0.940
- Interpretation: The model correctly predicted caravan insurance policy uptake for 94.0% of the cases in the test dataset.

Implications

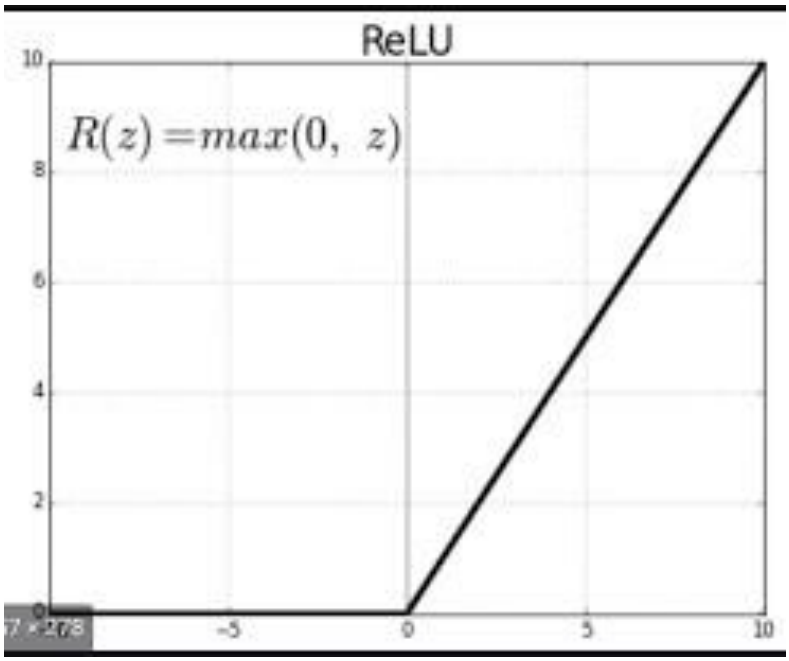
- Model Efficiency: High accuracy indicates strong predictive performance.
- Application: This model can be used to identify potential customers for caravan insurance policies.

Model 2 : Neural Network with full model

Input layer size : 85(No of variables)
Output layer size : 1 (Caravan)
Hidden Layer size : 40



Activation Function used ReLU



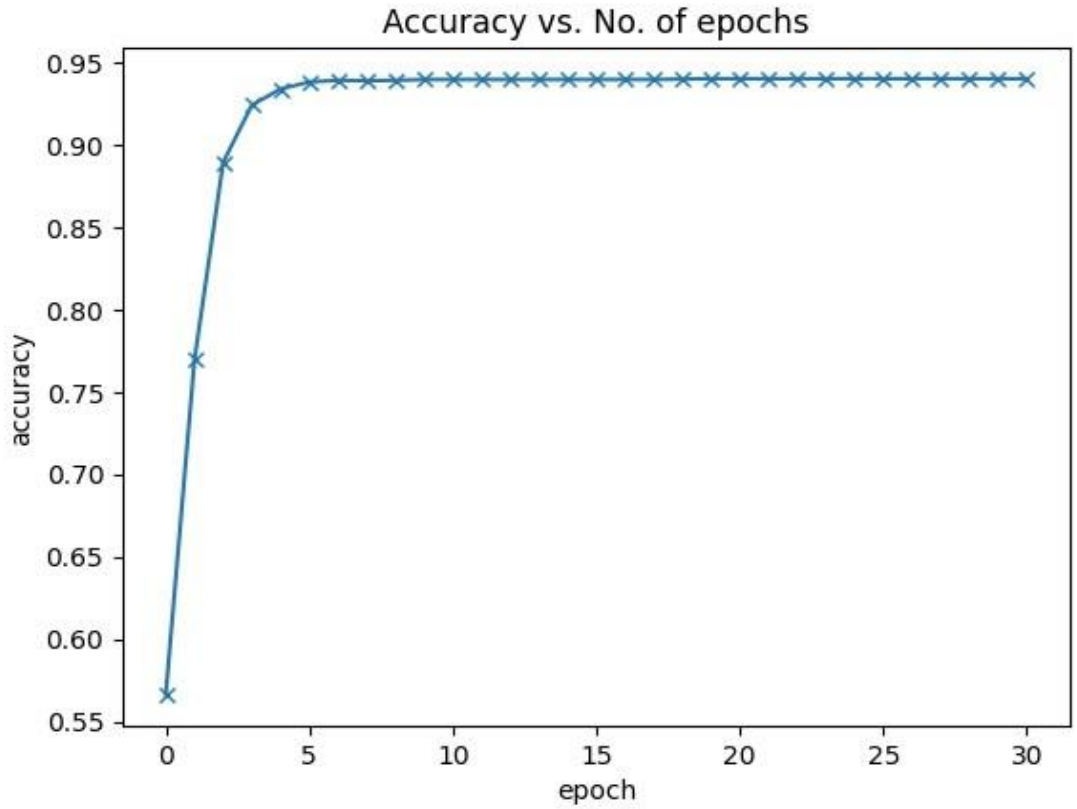
Loss function is:
binary_cross_entropy_with_logits

Optimizer Used for back Propagation :
Gradient Descent (SGD)

```
{'val_loss': 0.6912335157394409, 'val_acc': 0.5662500262260437}
```

```
history1 = fit(30, model, train_loader, val_loader)
```

```
Epoch [0], val_loss: 0.5888, val_acc: 0.7695
Epoch [1], val_loss: 0.5135, val_acc: 0.8895
Epoch [2], val_loss: 0.4572, val_acc: 0.9242
Epoch [3], val_loss: 0.4143, val_acc: 0.9340
Epoch [4], val_loss: 0.3809, val_acc: 0.9382
Epoch [5], val_loss: 0.3546, val_acc: 0.9393
Epoch [6], val_loss: 0.3336, val_acc: 0.9390
Epoch [7], val_loss: 0.3165, val_acc: 0.9393
Epoch [8], val_loss: 0.3025, val_acc: 0.9397
Epoch [9], val_loss: 0.2909, val_acc: 0.9397
Epoch [10], val_loss: 0.2812, val_acc: 0.9397
Epoch [11], val_loss: 0.2729, val_acc: 0.9397
Epoch [12], val_loss: 0.2659, val_acc: 0.9397
Epoch [13], val_loss: 0.2599, val_acc: 0.9397
Epoch [14], val_loss: 0.2547, val_acc: 0.9397
Epoch [15], val_loss: 0.2501, val_acc: 0.9397
Epoch [16], val_loss: 0.2462, val_acc: 0.9397
Epoch [17], val_loss: 0.2427, val_acc: 0.9400
Epoch [18], val_loss: 0.2396, val_acc: 0.9402
Epoch [19], val_loss: 0.2368, val_acc: 0.9402
Epoch [20], val_loss: 0.2344, val_acc: 0.9402
Epoch [21], val_loss: 0.2322, val_acc: 0.9402
Epoch [22], val_loss: 0.2302, val_acc: 0.9402
Epoch [23], val_loss: 0.2284, val_acc: 0.9402
Epoch [24], val_loss: 0.2268, val_acc: 0.9402
Epoch [25], val_loss: 0.2255, val_acc: 0.9402
Epoch [26], val_loss: 0.2242, val_acc: 0.9402
Epoch [27], val_loss: 0.2229, val_acc: 0.9402
Epoch [28], val_loss: 0.2218, val_acc: 0.9402
Epoch [29], val_loss: 0.2208, val_acc: 0.9402
```



Performance Metrics

Accuracy Score:

- Value: ~0.95
- Interpretation: The model correctly predicted caravan insurance policy uptake approximately ~95.0% of the cases in the test dataset.

Hypothesis Testing using Stepwise Regression:

What is Stepwise Regression?

Stepwise regression is a statistical method used for building regression models by adding or removing independent variables one at a time based on their statistical significance using a set criterion to see if it should remain in the model. Example of such a criterion is t value. It comes in two main flavors:

- Forward selection: Starts with no independent variables and adds the most statistically significant one at each step until no more improvements can be made.
- Backward elimination: Starts with all possible independent variables and removes the

least statistically significant one at each step until no more improvements can be made.

Benefits of Stepwise Regression:

- Reduces model complexity: By only keeping the most relevant independent variables, the model becomes simpler and easier to interpret.
- Improves model performance: By avoiding overfitting, the model can have better predictive power on unseen data.

Criteria to conclude a variable as a significant one is P-value should be less than 0.005(<5 percent) and absolute T-value should be highest.

Example:

- **Fit all 2-variable models: $Y = \beta_0 + \beta_1 X$?**

Y regressed on:	x1	x2	x3	x4	x5	x6	x7
b1	-0.001	0.128	0.03	0.013	0.049	0.556	0.019
t-statistic	-0.054	2.165	3.679	0.602	5.131	4.072	3.559
P value	0.957	0.035	0.001	0.55	0	0	0.004

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

Y regressed on:	x5+x1	x5+x2	x5+x3	x5+x4	x5+x6	x5+x7
b2	0.001	0.132	0.014	-0.032	0.358	0.007
t-statistic	1.422	2.817	1.68	-1.742	2.762	1.248
P value	0.162	0.007	0.099	0.087	0.008	0.2179

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

Y regressed on:	x5+x2+x1	x5+x2+x3	x5+x2+x4	x5+x2+x6	x5+x2+x7
b3	0.0014	0.0083	-0.0347	0.2582	0.0048
t-statistic	1.5004	0.9744	-2.0238	1.9001	0.8854
P value	0.1401	0.3347	0.0486	0.0634	0.3804

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

Y regressed on:	X5+X2+X4+X1	X5+X2+X4+X3	X5+X2+X4+X6	X5+X2+X4+X7
b4	0.0015	0.0155	0.2523	0.0051
t-statistic	1.7579	1.8066	1.9156	0.9664
P value	0.0853	0.0772	0.0615	0.3388

Using this step-wise algorithm and the above mentioned criteria on our dataset we concluded the following results.

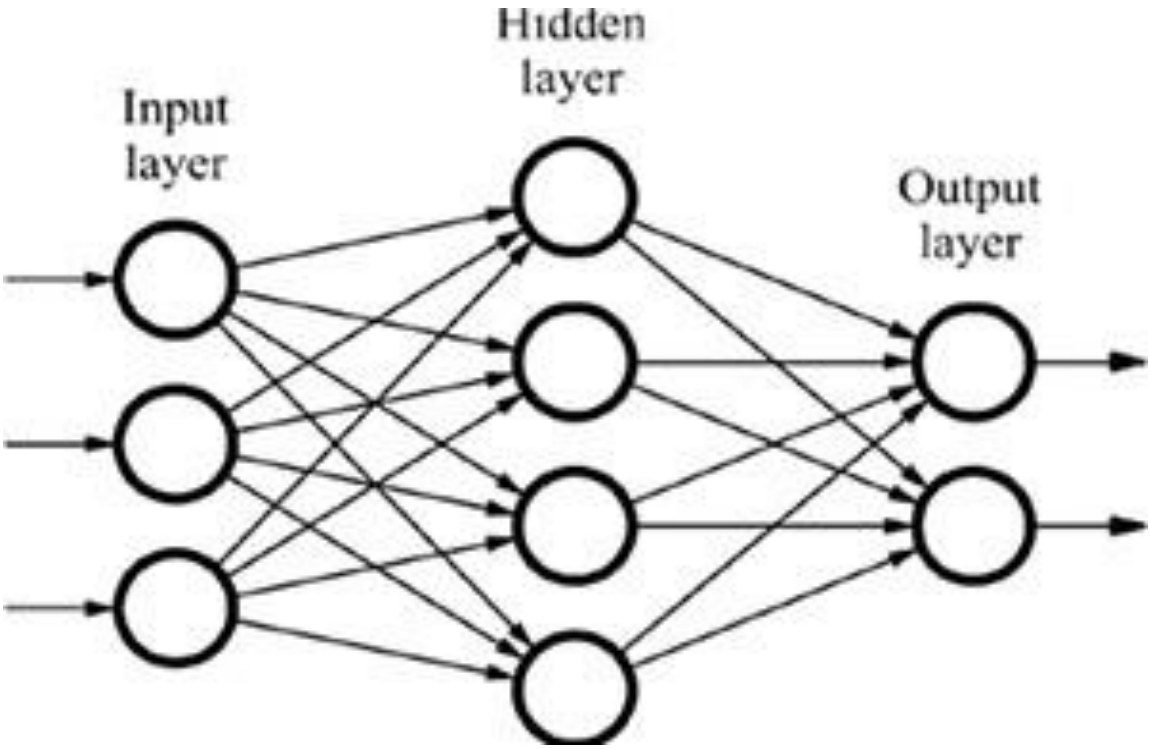
```
PPERSAUT                        with p-value 2.14684e-42
MKOOPKLA                        with p-value 1.36739e-21
PWAPART                         with p-value 3.66711e-15
APLEZIER                        with p-value 8.20766e-15
MOPLHOOG                        with p-value 4.25236e-06
PBRAND                          with p-value 3.92829e-06
MBERBOER                        with p-value 8.31838e-06
MRELGE                          with p-value 1.41977e-05
PWALAND                         with p-value 0.000361295
ABRAND                          with p-value 0.000937601
AZEILPL                         with p-value 0.00153041
MINK123M                        with p-value 0.00152554
PBYSTAND                        with p-value 0.00243579
PGEZONG                         with p-value 0.00485648
AGEZONG                         with p-value 0.00450709
MHHUUR                          with p-value 0.00630075
```

Resulting Features:
['PPERSAUT', 'MKOOPKLA', 'PWAPART', 'APLEZIER', 'MOPLHOOG', 'PBRAND', 'MBERBOER', 'MRELGE', 'PWALAND', 'ABRAND', 'AZEILPL', 'MINK123M', 'PBYSTAND', 'PGEZONG', 'AGEZONG', 'MHHUUR']

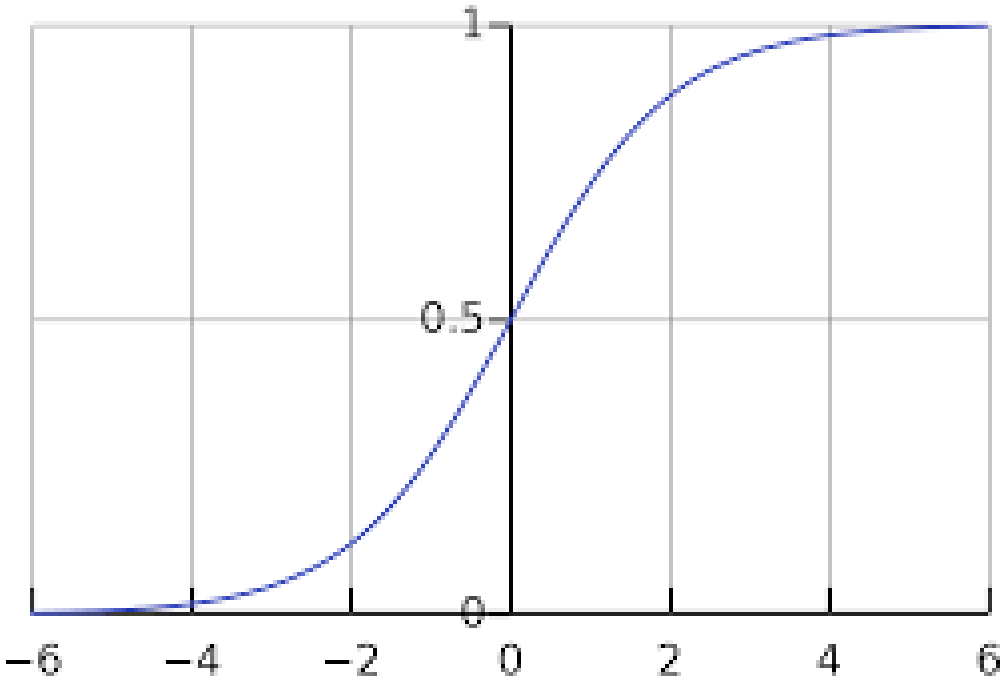
These 16 variables out of 85 are the most significant variables in our dataset according to our stepwise regression algorithm and the criteria.
So, with the stepwise regression model, we were able to reduce the variables from 85 to 16.
So, using only the important 16 variables given by Stepwise Regression, will now check the performance of neural network trained just with these 16 variables.

Model 3: Neural Network Trained with 16 variables from Hypothesis Testing

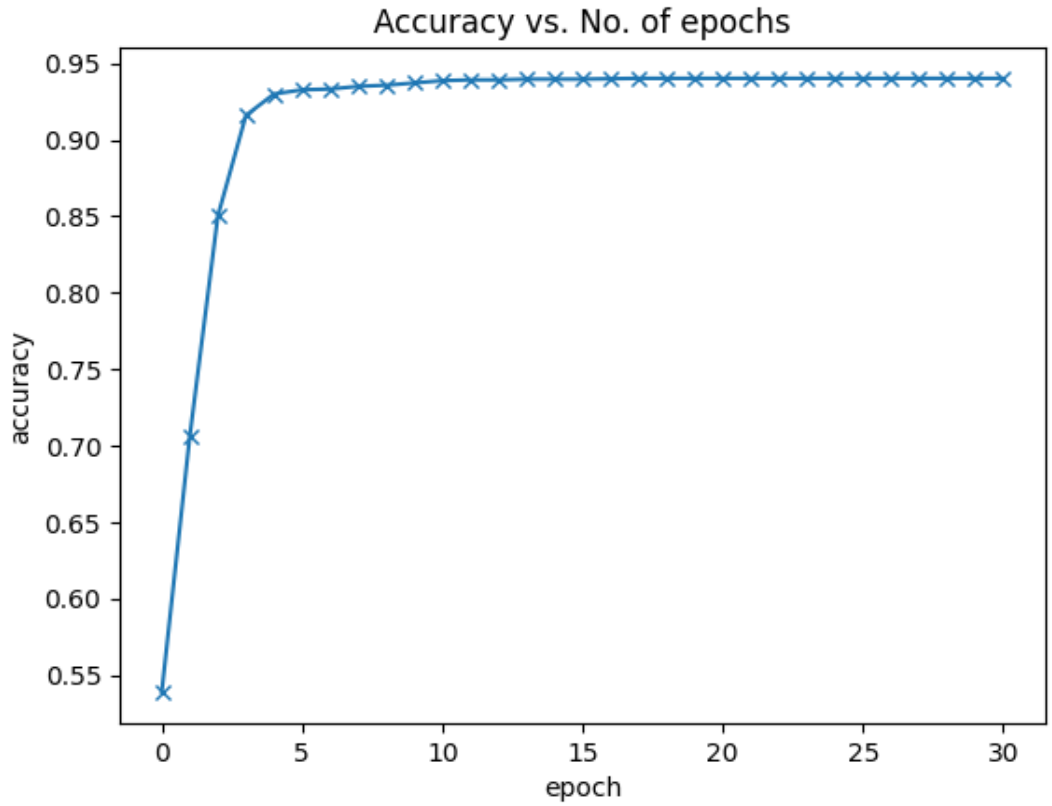
Input layer size : 16(No of variables)
Output layer size : 1 (Caravan)



Activation Function used SIGMOID



Loss function is:
binary_cross_entropy_with_logits
Optimizer Used for back Propagation :
Gradient Descent(SGD)




```
[ ]  result1 = evaluate(model, val_loader)
     result1

     {'val_loss': 0.7212085723876953, 'val_acc': 0.5390000343322754}

     history1 = fit(30, 0.001, model, train_loader, val_loader)

     Epoch [0], val_loss: 0.6137, val_acc: 0.7063
     Epoch [1], val_loss: 0.5330, val_acc: 0.8503
     Epoch [2], val_loss: 0.4721, val_acc: 0.9160
     Epoch [3], val_loss: 0.4255, val_acc: 0.9300
     Epoch [4], val_loss: 0.3894, val_acc: 0.9327
     Epoch [5], val_loss: 0.3610, val_acc: 0.9332
     Epoch [6], val_loss: 0.3384, val_acc: 0.9350
     Epoch [7], val_loss: 0.3201, val_acc: 0.9357
     Epoch [8], val_loss: 0.3052, val_acc: 0.9373
     Epoch [9], val_loss: 0.2928, val_acc: 0.9388
     Epoch [10], val_loss: 0.2825, val_acc: 0.9392
     Epoch [11], val_loss: 0.2738, val_acc: 0.9392
     Epoch [12], val_loss: 0.2665, val_acc: 0.9397
     Epoch [13], val_loss: 0.2601, val_acc: 0.9397
     Epoch [14], val_loss: 0.2547, val_acc: 0.9397
     Epoch [15], val_loss: 0.2500, val_acc: 0.9400
     Epoch [16], val_loss: 0.2458, val_acc: 0.9402
     Epoch [17], val_loss: 0.2422, val_acc: 0.9402
     Epoch [18], val_loss: 0.2390, val_acc: 0.9402
     Epoch [19], val_loss: 0.2362, val_acc: 0.9402
     Epoch [20], val_loss: 0.2337, val_acc: 0.9402
     Epoch [21], val_loss: 0.2314, val_acc: 0.9402
     Epoch [22], val_loss: 0.2294, val_acc: 0.9402
     Epoch [23], val_loss: 0.2276, val_acc: 0.9402
     Epoch [24], val_loss: 0.2259, val_acc: 0.9402
     Epoch [25], val_loss: 0.2244, val_acc: 0.9402
     Epoch [26], val_loss: 0.2231, val_acc: 0.9402
     Epoch [27], val_loss: 0.2218, val_acc: 0.9402
     Epoch [28], val_loss: 0.2207, val_acc: 0.9402
     Epoch [29], val_loss: 0.2197, val_acc: 0.9402
```

We got 94 percent accuracy same as when we have trained with all 85 variables so we can conclude that these 16 variables that we got from stepwise regression.
We can conclude out of 85 variables 16 are important and rest all variables are redundant.

- **Model 3: Random Forest**

  What is Random Forest?

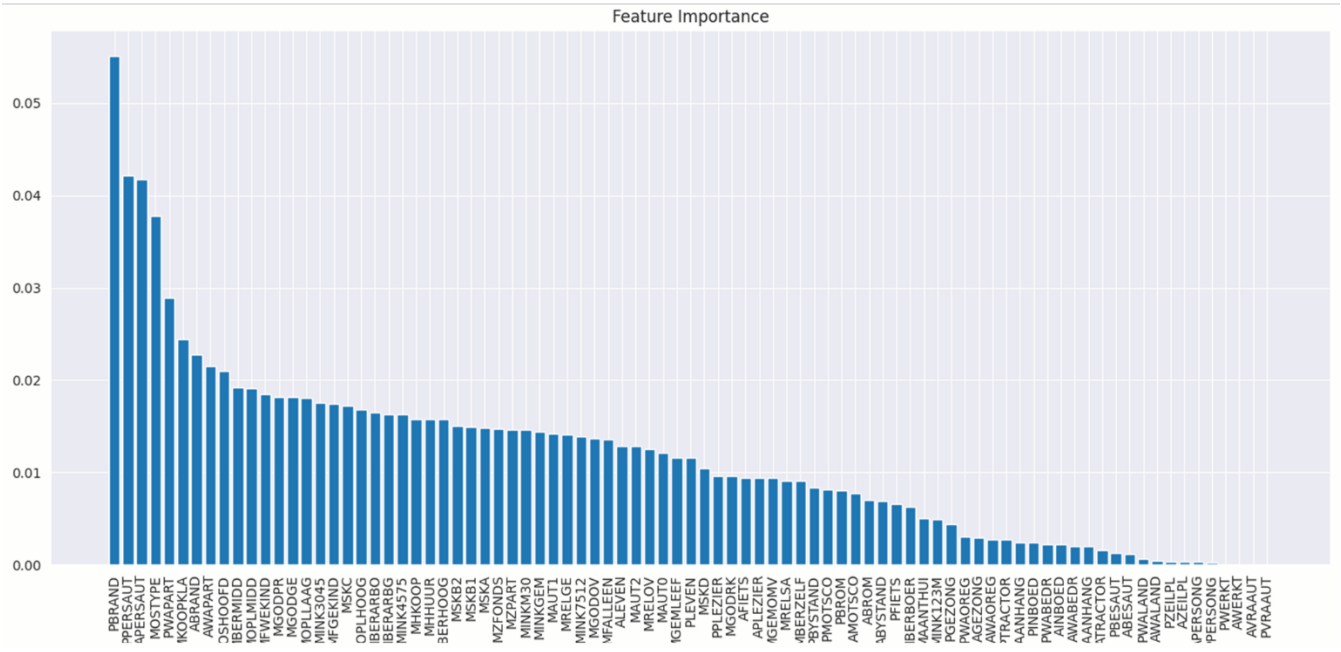
  Random Forest is an ensemble learning method for classification and regression. It constructs multiple decision trees during training and outputs the mode of classes (classification) or mean prediction (regression) from individual trees.

  How Does Random Forest Work?
  Bootstrapped Sampling: Builds trees by sampling training data with replacement, training each tree on a different data subset.
  Feature Randomness: Considers a random subset of features at each decision tree node, introducing diversity.
  Voting (Classification) or Averaging (Regression): Trees "vote" for classes in classification or contribute predictions for regression. The majority class or average prediction is the final output.
  Key Advantages:
  High Accuracy: Often achieves high prediction accuracy.
  Reduced Overfitting: Ensemble approach and randomness enhance generalization.
  Reasons for using random forest as a performance reference:
  While powerful, Random Forest models are often considered less interpretable compared to linear models like stepwise regression. The ensemble nature of Random Forest makes it challenging to trace the decision-making process. It gives Importance of feature in magnitude but leaves a gap interpreting in what way feature affects the prediction.
  Might not perform as well with very small datasets due to the ensemble approach. Training multiple decision trees can be computationally expensive, especially for large datasets or a high number of trees.
  Data Set is split into test and training data after dropping the label column. Train Data has 7000 data points to ensure good ensemble approach. With good amount of data wide variety of subsets can be generated decreasing bias.
  The number of trees, represented by the `n_estimators` parameter, is crucial in a Random Forest model, affecting overall performance. Finding the optimal number involves balancing computational efficiency and accuracy. Our approach is to empirically test different `n_estimators` values, weighing computational cost against accuracy, to determine the best choice.

  RF Results:

  - Accuracy: 0.931
  - Top 10 most important features: PBRAND, PPERSAUT, APERSAUT, MOSTYPE, PWAPART, MKOOPKLA, ABRAND, AWAPART, MOSHOOFD, MBERMIDD
  - Feature selection:
    Threshold on feature importance scores (importance score > 0.02): 16 features

- Recursive feature elimination (RFE): variable number of features depending on the desired performance/speed trade-off



Conclusion

From the Random Forest results, we can interpret that the factors: “Contribution fire policies, Contribution car policies, Number of car policies, Customer Subtype, Contribution private third-party insurance, purchasing power class, Number of fire policies, Number of private third-party insurance, Customer main type, Middle management” highly influence the chances a person being potential customer.

But from random forest results we cannot interpret if the factors negatively/ positively contribute to the chances of potential customer.

These understandings are used as a base to test our step wise regression model accuracy in prediction after dropping variables that contribute the least.

Correlation Matrix analysis identifies features associated with the likelihood of purchasing caravan insurance.

Increased Likelihood:

Family Status: Married or partnered, with children.

Financial: Higher income

Caravan Ownership: Owning a more expensive caravan.

Demographics: Older age

Decreased Likelihood:

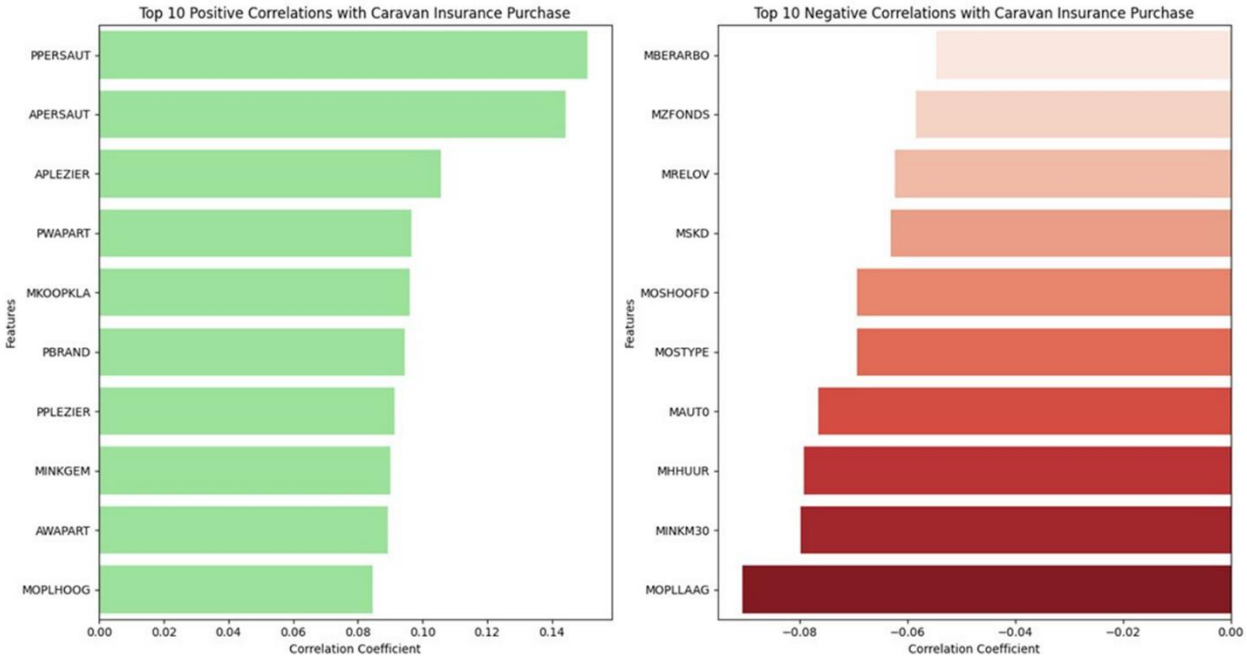
Financial: Lower income

Caravan Ownership: Renting a caravan, owning a less expensive caravan.

Demographics: Younger age

These correlations suggest potential customer segments for targeted marketing campaigns.

Remember, correlation doesn't imply causation, but these insights can help tailor marketing efforts to specific customer segments.



From Step-Wise regression analysis, Out of the 85 variables initially considered, only 16 are needed to predict whether a customer takes insurance. This means that the other 69 variables are not significant for prediction purposes and can be excluded. This significantly reduces the computational resources and other resources needed for analysis, making the process more efficient and cost-effective.

Contribution car policies	Married
Contribution family accidents insurance policies	High level education
Contribution fire policies	Farmer
Contribution social security insurance policies	Rented house
Number of family accidents insurance policies	Income >123.000
Number of fire policies	Purchasing power class
Number of surfboard policies	Contribution private third party insurance
Number of boat policies	Contribution third party insurance (agriculture)

9 of these variables also appear in the top 20 most correlated variables with 'caravan' in the correlation matrix and appears in the top 20 important features in Random Forest Results. This overlap underscores the high significance of these 9 variables. Given their strong correlation and the validation from two distinct analytical methods, these variables should be considered especially critical. In future data collection efforts, prioritizing these variables could provide more insightful and reliable analyses.

Number of boat policies
Rented house
Purchasing power class
High level education
Contribution fire policies
Contribution car policies
Contribution private third party insurance

In conclusion, our project focuses on identifying vital variables crucial for precise predictions of target customers in the context of caravan insurance. By emphasizing the significance of these key factors, our approach is geared towards minimizing overall business costs, cutting down on computational expenses, and optimizing resource utilization. This strategic prioritization of impactful variables not only enhances marketing performance but also holds the promise of driving increased customer conversions. Through this targeted approach, we aim to achieve optimal efficiency in marketing efforts while ensuring a prudent allocation of resources in the realm of caravan insurance.

References:

https://archive.ics.uci.edu/static/public/125/insurance+company+benchmark+coil+2000.zip

https://jovian.com/aakashns/03-logistic-regression