Report - Assignment 2

Train_C

Contains the Procedure followed to complete the assignment. Also contains the Hyperparameters chosen based on Experimentation.

The First Step is choosing if we want to consider a feature as Discrete, whose probability estimation is done using frequency of occurrences, or if we want to consider it as Continuous, whose probability estimation is done assuming it as a Gaussian Distribution.

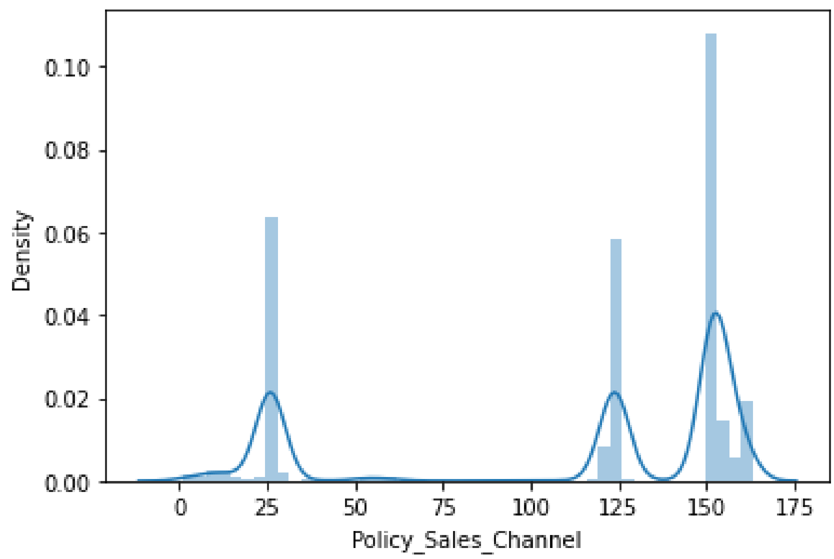
Feature Type

Discrete Features:

```
1) Gender
2) Driving_License
3) Previously_Insured
4) Vehicle_Age
5) Vehicle_Damage
6) Policy_Sales_Channel
```

Reason behind taking Policy_Sales_Channel as Discrete:

Plotting the distribution of Policy_Sales_Channel

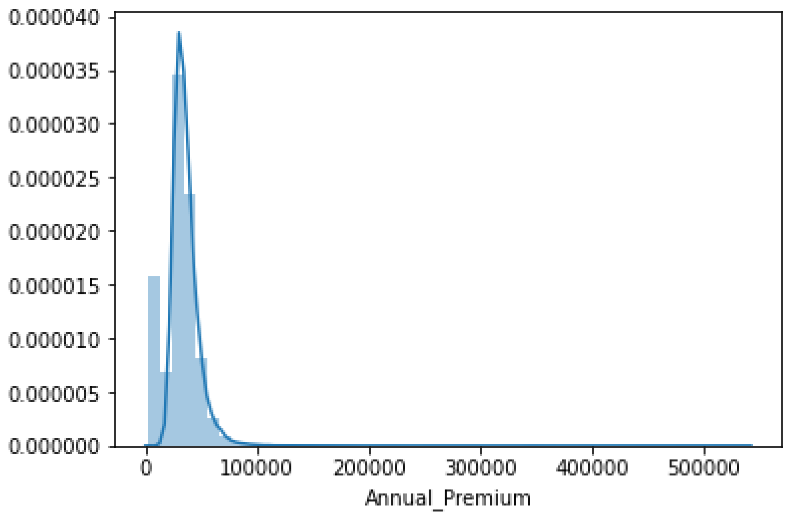


- The above plot shows that the Policy_Sales_Channel looks less like a Gaussian Distribution and more like a Discrete Variable with a high number of Data Points confined to 3 regions.
- Hence we Discretised this feature into 3 classes: <50,   >50 and < 150,   >150
  This gave us the best results

Continuous Features:

```
1) Age
2) Region_Code
3) Annual_Premium
4) Vintage
```

These features are considered as Continuous variable.



Probability Computation of Discrete vs Continuous:

Discrete:

The probability computation for Discrete classes is simply the frequency of occurrence in a particular class

P(Ai/Ci) = $\frac{\#\ of\ occurences\ of\ Ai\ \ in\ the\ examples\ having\ Ci\ target}{\#\ of\ total\ exaples\ with\ target\ Ci}$

In our code we separately compute all these probabilities and store it in a dictionary.

We note that the above computation can lead to a Zero Probability which will make the whole Probability Function zero. Hence, we use the m-estimator method to do zero correction.

M estimator method:

m-estimate of probability:

$$\frac{n_c + mp}{n + m}$$

- Here m is the variable hyperparameter and p is the Prior estimate of the attribute. We assume all values of the attribute are equally probable and hence p = 1/(No. of unique values of attribute).

- We note that Varying the m values considerably changed the accuracy:
- The extent of variation is discussed in the results page.

Continuous Variable:

- To compute the probability of a Continuous variable, we first assume that the variable is approximately Gaussian in nature
- We calculate the mean and the standard deviation of a particular feature corresponding to class 0 and class 1 separately.
- With the mean and stdev the Gaussian Probability can be computed as follows:

P(Xi/ Ci) = $\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2 / 2\sigma^2}$ where σ is the stdev of a particular attribute belonging to c lass Ci

Once these Probabilities are computed, we simply use Bayes Theorem to get the probability of Class given data.

P(Ci/ Ai) = P(a1*a2*a3……/Ci)*P(Ci)/P(Ai)

The marginal probability is ignored during the computation. The class that gives the Highest numerator value is considered the correct class.

Note: **We assume the features/attributes are conditionally independent. Hence**

P(a1*a2……/Ci) = P(a1/Ci)*P(a2/Ci)……..

We appropriately take the Gaussian Probability if it’s a continuous feature or take the Frequency based probability if it’s a discrete variable,

With this computation we make the predictions.

PART 3:

Deleting the outlier features:

- We were told that a data point should be deleted if more than 50% of its features are outliers:
- Outlier means the value is greater than mean + 3*stdev or less than mean – 3*stdev
- In our case we are hardly getting deletion of data points, so we modified it such that a data point gets deleted if more than 30 % of its features are outliers.

Feature Selection:

Backward Feature selection was performed and we got 7 features as the best selected features out of 10 in total

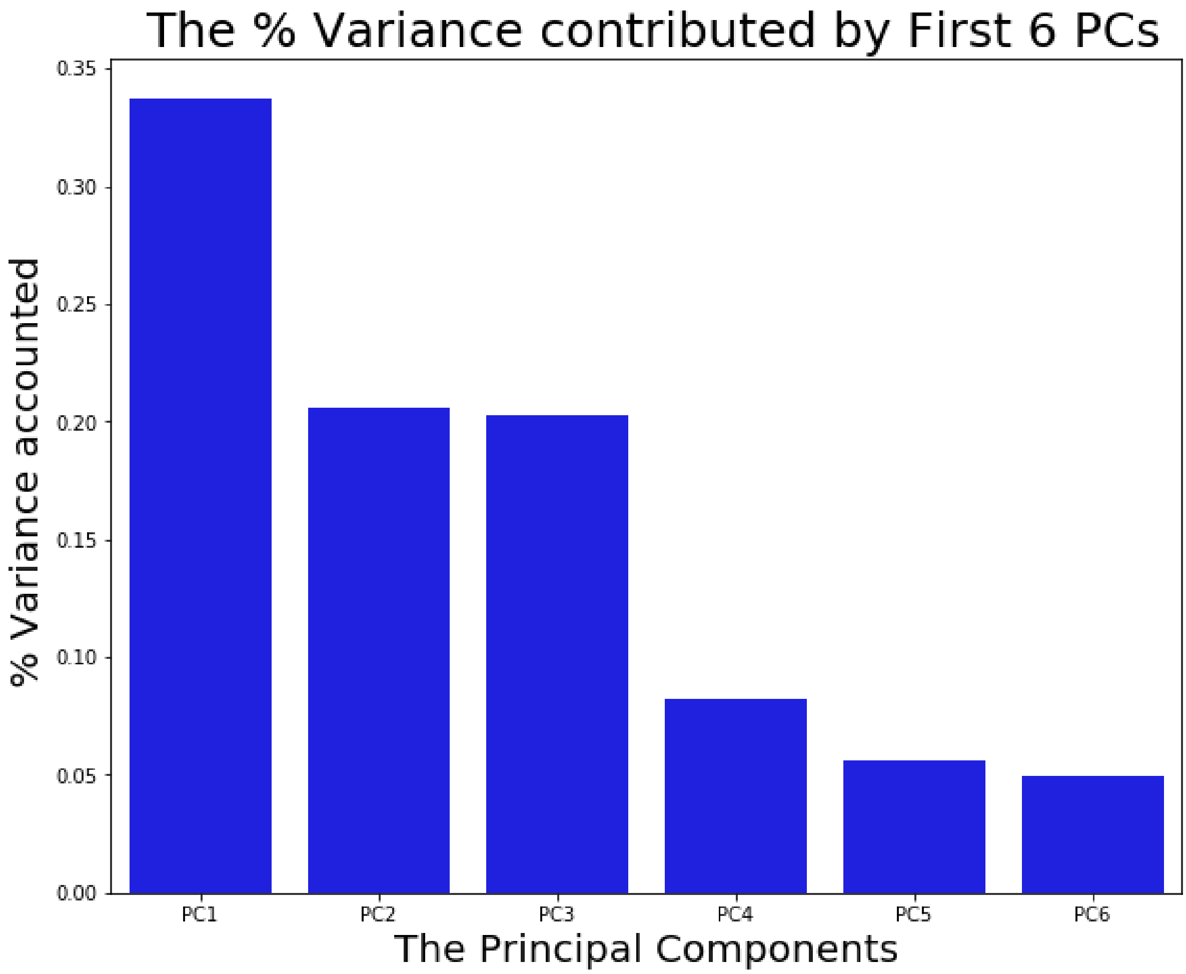
Selected Features =
```
['Gender',
 'Age',
 'Driving_License',
 'Region_Code',
 'Previously_Insured',
 'Policy_Sales_Channel',
 'Vintage']
```

Part 2

Principal Component Analysis

- The data had to be normalised before applying PCA, but we had already normalised the values at the starting
- We had to preserve 95% of the variance.
- One Computing we saw the First 6 Principal Components retained 94% of the Variance, hence the following computation was done with the first 6 PC taken as Continuous Variables

The Variance Captured by PCs



Principal Space Visualisation: