## Drug store data set

* **Objective:** The Drug store plans to launch a pregnancy campaign for its Loyalty card members and would to do a personalized targeted campaign. They would like to predict based on past transactions likelihood of preparing for pregnancy in a household and take the following actions:
  + talk to the customer at POS or
  + send a targeted email or
  + invite for a pregnancy campaign
* **Background:** Data for 2000 Loyalty Card members of a popular drug store is given
* **Variables:** several demographic and buying behavior variables are given for our understanding, including whether there was a Pregnancy Yes or No
* **QUESTIONS TO ANSWER:**

1. Which is better Regression or Classification to solve such a problem? [compare results]
2. What variables should you consider to predict if a household is facing Pregnancy or No?
3. How did you draft a Test vs Validation data set? Should you follow a standard? Discuss for both Regression and Classification in detail.

**Answers:**

1. Classification is better than Regression because of the following results below are the classification report and the confusion matrix of each.

Regression model:

In regression I used two types of regression they are Linear Regression and Logistic Regression.

Linear Regression:

In this model the accuracy which I got was not satisfying. The accuracy for this model obtained was 42% which is very low. So, for regression analysis, for this kind of problem that is having binary values in the variable as a target variables Logistic regression would be better and would give better results than Linear Regression.

Logistic Regression:

Considering the Logistic Regression model the accuracy obtained for this model is 68% below are the tables of confusion matrix and the classification report which would provide how the accuracy is calculated and also the factors which were considered for calculating the result.

From the regression analysis I can conclude that, the results obtained weren’t satisfying so it would be better taking up the classification model for the analysis.

To analyze these model we use two tables they are:

Confusion matrix:

A confusion matrix is a table that is often used to **describe the performance of a classification model** (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.

* There are two possible predicted classes: "yes" and "no". If we were predicting the presence of a disease.
* Let's now define the most basic terms, which are whole numbers.
* **true positives (TP):** These are cases in which we predicted yes, and they are actually yes.
* **true negatives (TN):** We predicted no, and they are actually no.
* **false positives (FP):** We predicted yes, but they don't actually they are no. Also known as a "Type I error."
* **false negatives (FN):** We predicted no, but they are actually yes. (Also known as a "Type II error.")

This is a list of rates that are often computed from a confusion matrix for a binary classifier:

* **Accuracy:** Overall, how often is the classifier correct?
  + (TP+TN)/total
* **Misclassification Rate:** Overall, how often is it wrong?
  + (FP+FN)/total
  + equivalent to 1 minus Accuracy
  + also known as "Error Rate"
* **True Positive Rate:** When it's actually yes, how often does it predict yes?
  + TP/actual yes , also known as "Sensitivity" or "Recall"
* **False Positive Rate:** When it's actually no, how often does it predict yes?
  + FP/actual no.
* **Specificity:** When it's actually no, how often does it predict no?
  + TN/actual no ,equivalent to 1 minus False Positive Rate
* **Precision:** When it predicts yes, how often is it correct?
  + TP/predicted yes.
* **Prevalence:** How often does the yes condition actually occur in our sample?
  + actual yes/total .

Classification Report:

* Compute precision, recall, F-measure and support for each class
* The precision is the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative.
* The recall is the ratio tp / (tp + fn) where tp is the number of true positives and fn the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples.
* The F-beta score can be interpreted as a weighted harmonic mean of the precision and recall, where an F-beta score reaches its best value at 1 and worst score at 0.
* The F-beta score weights recall more than precision by a factor of beta. beta == 1.0 means recall and precision are equally important.
* The support is the number of occurrences of each class in y true.

|  |  |  |
| --- | --- | --- |
|  | Predicted No | Predicted yes |
| Actual No: | TN = 138 | FP = 70 |
| Actual Yes: | FN = 57 | TP = 135 |

Confusion matrix for Logistic Regression:

Classification report for Logistic Regression:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Result | Precision | recall | f1-score | support |
| No | 0.70 | 0.66 | 0.68 | 208 |
| Yes | 0.66 | 0.70 | 0.68 | 192 |
| average / total | 0.68 | 0.68 | 0.68 | 400 |

Classification Model:

For classification model I have considered the Decision Tree Classifier. Compared to Regression model this model gave better results. The accuracy obtained for this model is 96%. Below are the Classification Report and confusion matrix tables which will help in calculating the accuracy and also show the other factors which were considered while calculating the accuracy of the model.

Confusion matrix for Decision Tree Classifier.

|  |  |  |
| --- | --- | --- |
|  | Predicted No | Predicted yes |
| Actual No: | TN = 201 | FP = 7 |
| Actual Yes: | FN = 11 | TP = 181 |

Classification report for Decision Tree Classifier.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Result | Precision | recall | f1-score | support |
| No | 0.95 | 0.97 | 0.96 | 208 |
| Yes | 0.96 | 0.94 | 0.95 | 192 |
| average / total | 0.96 | 0.95 | 0.95 | 400 |

From the above reports on regression and classification I can conclude that classification model is giving better results than regression model.

2. The variables which were considered as features for the prediction are:

* Implied Gender
* Age
* Discount Prg
* Celebrated Valentines
* Pregnancy Test
* Birth Control
* Feminine Hygiene
* Thanks Giving Shopping
* Number Shopping trips
* Folic Acid
* Prenatal Vitamins
* Prenatal Yoga
* Body Pillow
* Ginger Ale
* Nausea tablets
* Stopped buying ciggies
* Cigarettes
* Smoking Cessation
* Stopped buying wine
* Wine
* Maternity Clothes

Except Loyalty card number and Residence type columns all the other columns were considered as features for predicting whether the person in pregnant or not.

In above mentioned columns except age rest all the other variables are categorical values having only yes or no(Only Implied\_gender had Male female and unknown as their values) as their values so these variables were converted to binary values having ones and zeros as their values. It was converted using get\_dummies function present in pandas package. After converting these variables, the variables are preprocessed using StandardScaler function present in sklearn.preprocessing package. After the data is preprocessed dimensionality reduction process is done on the features, this process is done by using the PCA function present in the sklearn.decompostion package. The following number of dimensions were taken up in each model are as follows:

Linear Regression: 50

Logistic Regression: 30

Decision Tree Classifier: 30

So in Linear Regression model, 50 dimensions were taken, in Logistic Regression 30 dimensions and for Decision Tree Classifier 30 dimensions. After the dimensionality reduction the dataset will be divided into training data and testing and then these datasets are sent for training the models defined.

3. Test vs validation data set for this problem for both Regression and Classification were taken randomly. This was done by using train\_test\_split function present in sklearn.crossvalidation package. 80% of the data were taken as training and 20% for testing (validation).

The model was trained by taking 80% of the dataset and then these models were used to predict the values of the test data for validation so as to check the accuracy for each model. So after training the model these models can be used to check for the target variable (PREGNANT) that is to check whether the person is pregnant or not. These models can be also used by the store just by giving the values to the variables (columns that are taken as features) as input and then they can check whether the person is pregnant or not by using these any of the models.