Blood Donation Data: Training Set

About the data:

1. There are no missing values.
2. There are 4 independent variables:
3. Months\_since\_Last\_Donation
4. Number\_of\_Donations
5. Total\_Volume\_Donated\_cc
6. Months\_since\_First\_Donation

**Exploratory Data Analysis:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sr.No | Months\_since\_Last\_Donation | Number\_of\_Donations | Total\_Volume\_Donated\_cc | Months\_since\_First\_Donation |
| 1 | Not normal, multimodal, positively skewed | Not normal, Positively skewed | Not normal, Positively Skewed | Not normal,  Kurtosis present |
| 2 | Outliers present | Outliers present | Outliers present | No outliers |
| 3 | Homogenity of variance violated | Homogenity of variance violated | Homogenity of variance violated | Homogenity of variance NOT violated. |
| 4 | Linearity assumption holds. | Linearity assumption violated. | Linearity assumption violated. | Linearity assumption holds. |

3 assumptions are important for logistic regression:

1. Linearity -> violated by one predictor.
2. Independence of errors
3. Number\_of\_Donations and Total\_Volume\_Donated\_cc are highly correlated.
4. Multicollinearity. -> not violated. (if u use only one of the above 2 variables).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Attempt  No. | About the data | Algorithm | Score | Comments |
| 1 | Used all cases. | Logistic Regression | 0.4457 | Residual Deviance = 556.61 |
| 2 | Used the "Total\_Volume\_Donated\_cc" column instead of "Number\_of\_Donations" column | Logistic Regression | 0.4457 |  |
| 3 | Used product of "Total\_Volume\_Donated\_cc" and "Number\_of\_Donations" columns | Logistic Regression | No Submission | Residual deviance increased to 573.65 |
| Attempt  No. | **About the data** | **Algorithm** | **Score** | **Comments** |
| 4 | Applied log transformation to all predictors in training set and test set. | Logistic Regression | 0.4654 |  |
| 5 | Used all cases and the actual dataset. | Naïve Bayes | 0.6829 |  |
| 6 | Used all cases and the actual dataset. | Support vector machine | 0.4769 |  |
| 7 | Applied log transformation to all predictors in training set and test set. | Support vector machine | 0.4663 |  |
| 8 | Split the training data in to training set and test set. Removed outliers. | Logistic regression | 0.4450 |  |
| 9 | Removed outliers and applied log transformation to predictors | Logistic regression | No Submission | On the test data we got Logloss = 0.535, which is same as the logloss we obtained without applying log transformation. |
| 10 | Calculated residual statistics and influential cases. Identified and removed case which had highest leverage value. | Logistic regression | 0.4449 | On the test data we got Logloss = 0.5349, which is a very slight improvement from the previous case. |
| 11 | Log Transformed only “Months\_since\_Last\_Donation” and “Number\_of\_Donations” predictor columsn | Logistic Regression | 0.4478 | On the test data, we got a lot loss of 0.5311, which is an improvement over previous attempt. |
| 12 | Performed Bootstrapping with logistic regression | Logistic Regression | 0.4546 | On the test data, we got a lot loss of 0.5142593, |