**Key insights from data analysis:**

* Given data set is an imbalanced data with 77% positive class and 23% negative class.

A graph of a loan status

Description automatically generated with medium confidence

* Found credit history has direct impact on loan status. If a person has credit history, the probability of getting a loan is more.
* Similarly, if a person completed graduation, he has higher probability of getting a loan.
* The chance of getting a loan is high If a person doesn’t have any dependents.
* Above 90% opted for a loan term of 360 months.

A group of graphs with numbers

Description automatically generated with medium confidence

**Modelling Part:**

* Handled outliers by capping them at 5 and 95 percentiles.
* Applied SMOTE with 0.7 as ratio between number of samples of minority class & majority class.
* StratifiedKFold is used while doing cross validation.

**Performance Metrics:**

* Since we are interested in less number of false positives, precision is the desired metric.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Name/Metric | DecisionTree Classifier | RandomForest Classifier | LogisticRegression | KNeighbors Classifier |
| Cross Validation Accuracy | 0.65 (+/- 0.04) | 0.73 (+/- 0.03) | 0.76 (+/- 0.04) | 0.71 (+/- 0.05) |
| Cross Validation Precision | 0.74 (+/- 0.03) | 0.78 (+/- 0.03) | **0.79 (+/- 0.03)** | 0.75 (+/- 0.03) |
| Cross Validation roc\_auc | 0.61 (+/- 0.07) | 0.72 (+/- 0.04) | 0.75 (+/- 0.06) | 0.64 (+/- 0.04) |

* Logistic Regression has given highest precision among different classifiers.