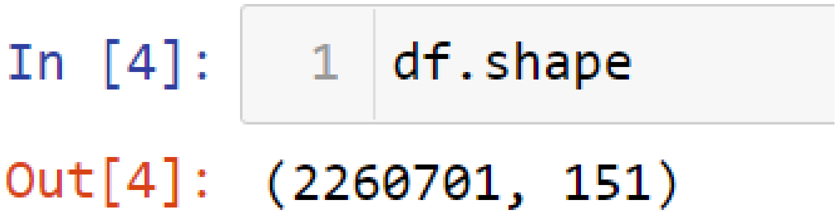
Loan default Classification using Logistic regression

Introduction

- **Dataset Source**: https://www.kaggle.com/wordsforthewise/lending-club
- **Lending Club** is the largest online loan marketplace, facilitating variety of loans.
- Dataset contain complete loan data for all loans issued through the 2007-2018.
- **Task**: Classify whether loan will be fully paid or charged off i.e. Binary Classification.

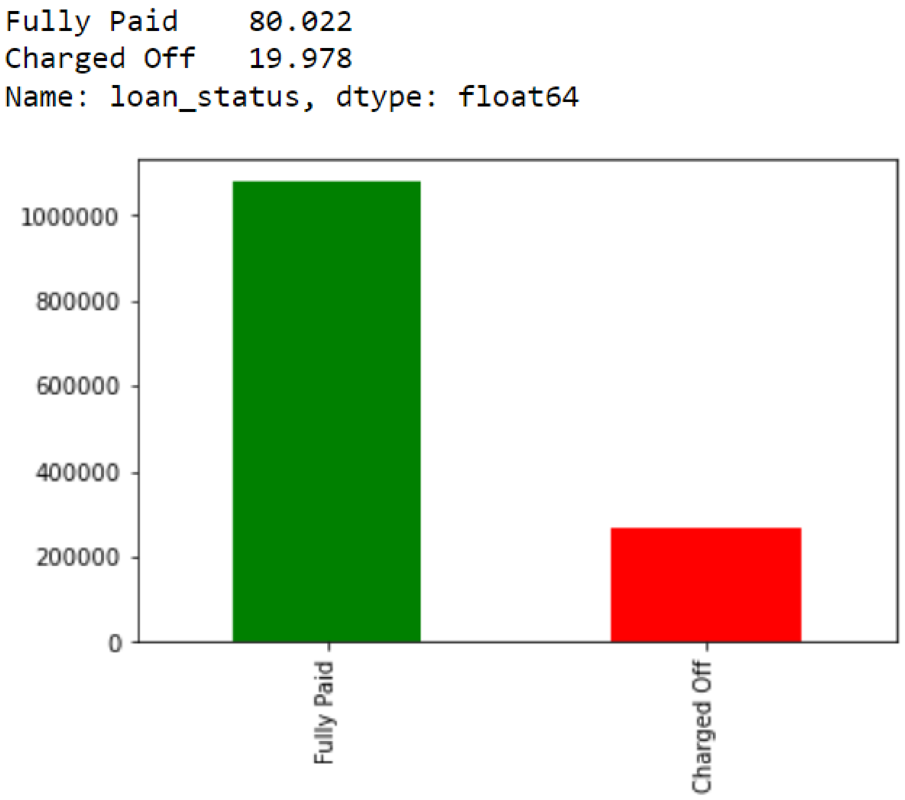


```
In [4]:  df.shape

Out[4]: (2260701, 151)
```

Challenges acts as a Motivation

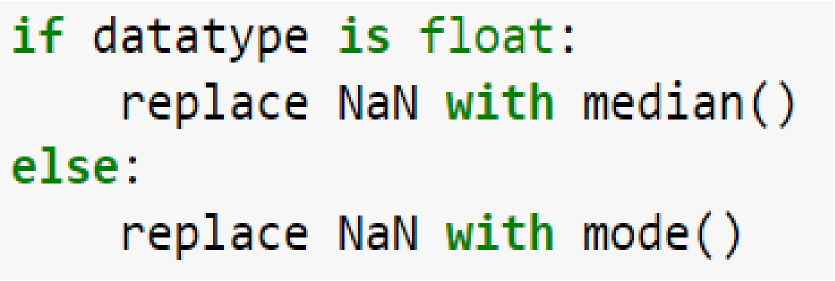
- Almost all columns having Missing values (i.e. NaN)
- Heavily imbalanced data: Fully Paid-80%, Charged off-20%
- Lots of categorical features.



```
Fully Paid     80.022
Charged Off    19.978
Name: loan_status, dtype: float64
```

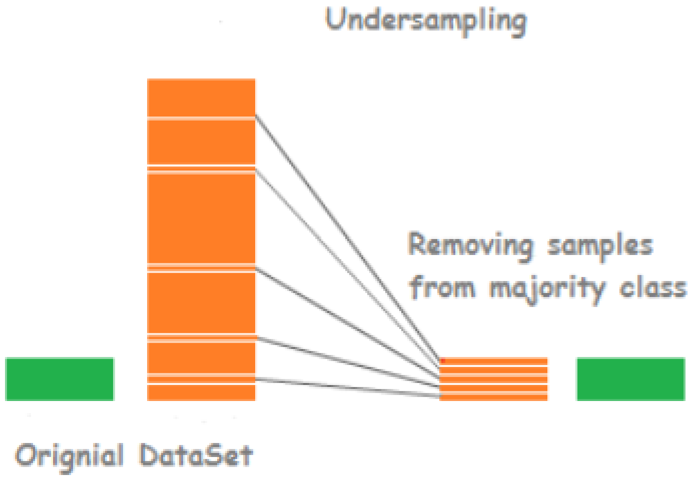
Preprocessing

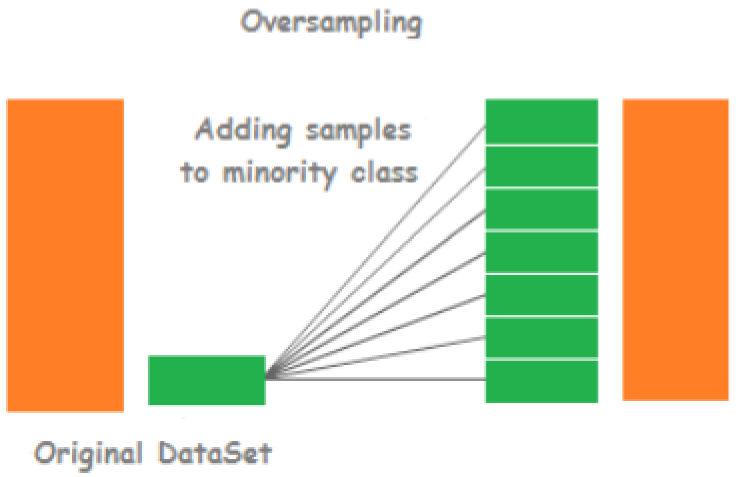
- Delete feature column having Missing values more than 30%
- Drop feature column having correlation value greater than 0.98
- Delete other unnecessary columns like 'id', 'emp_title', 'issue_d', 'pymnt_plan', 'url', 'title', 'zip_code', 'addr_state',.. etc.
- Consider only Fully paid/Charged off data
- Handling Missing data:



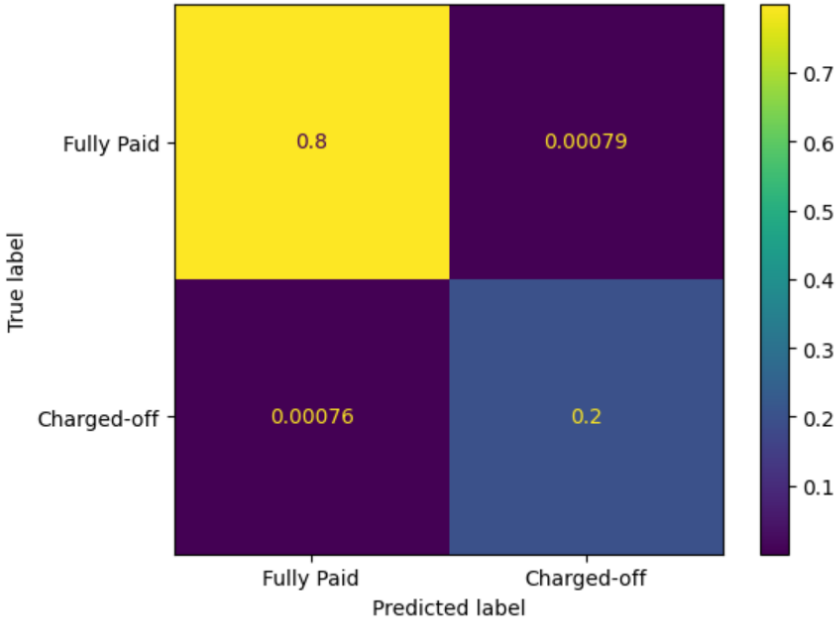
```
if datatype is float:
    replace NaN with median()
else:
    replace NaN with mode()
```

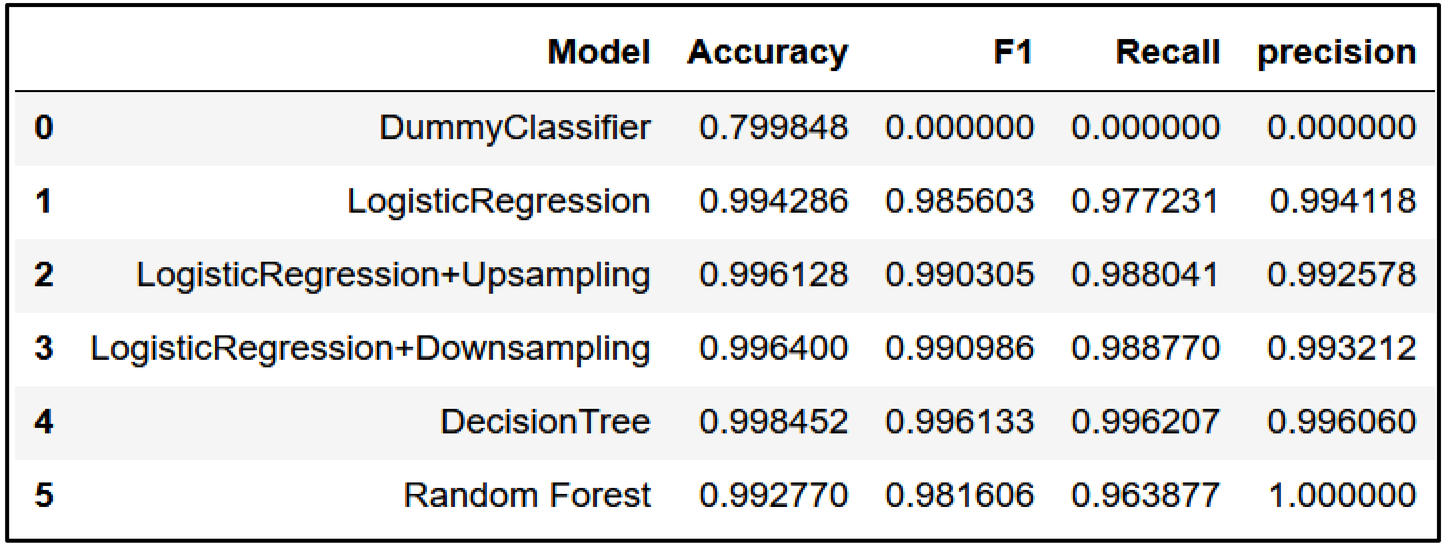
- Convert categorical features to numerical values
- Up/Down Sampling applied to balance the data





Results





	Model	Accuracy	F1	Recall	precision
0	DummyClassifier	0.799848	0.000000	0.000000	0.000000
1	LogisticRegression	0.994286	0.985603	0.977231	0.994118
2	LogisticRegression+Upsampling	0.996128	0.990305	0.988041	0.992578
3	LogisticRegression+Downsampling	0.996400	0.990986	0.988770	0.993212
4	DecisionTree	0.998452	0.996133	0.996207	0.996060
5	Random Forest	0.992770	0.981606	0.963877	1.000000

Decision Tree is the best model