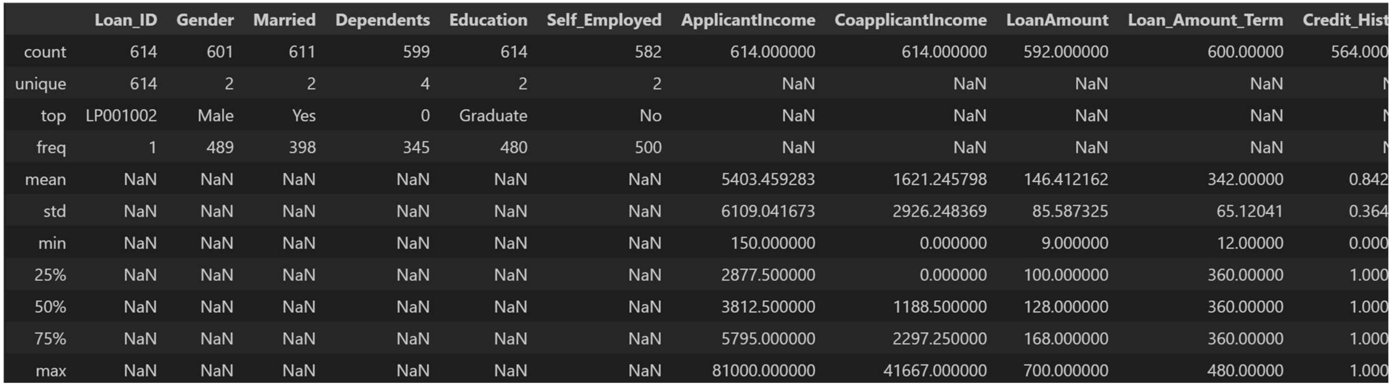
Data Collection and Preprocessing Phase

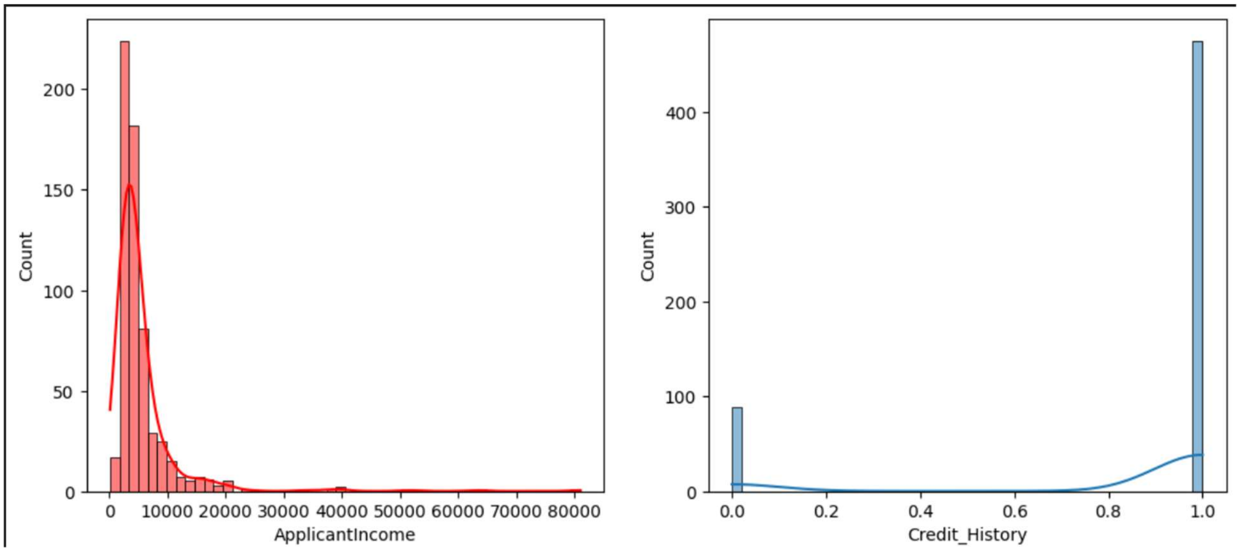
Date	19 June 2025
Team ID	SWTID1750050475
Project Title	SmartLender - Applicant Credibility Prediction for Loan Approval
Maximum Marks	6 Marks

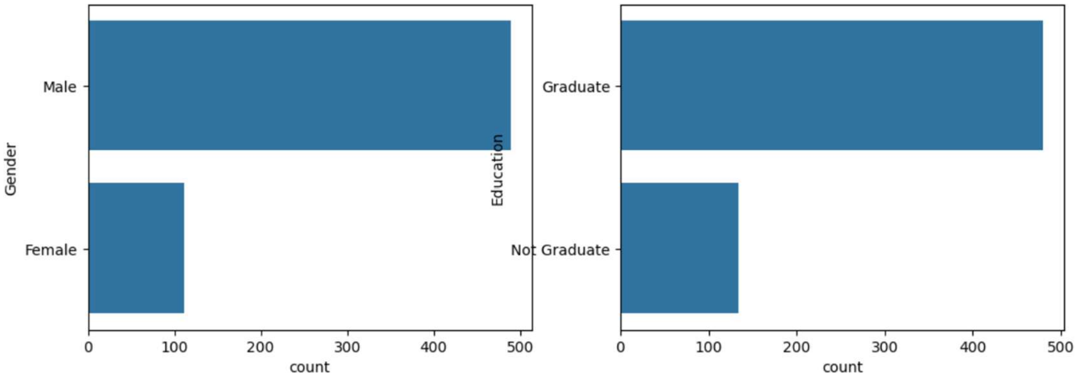
Data Exploration and Preprocessing Report

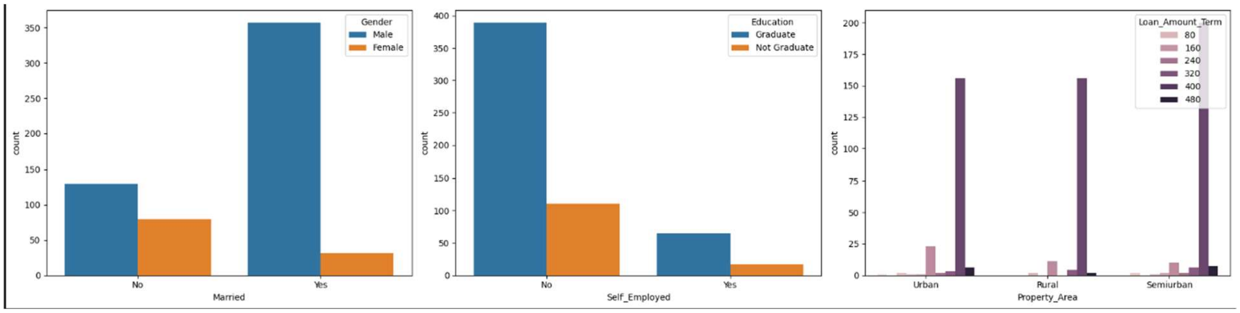
Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

Section	Description
Data Overview	Dimension: 614 rows × 13 columns Descriptive statistics: 

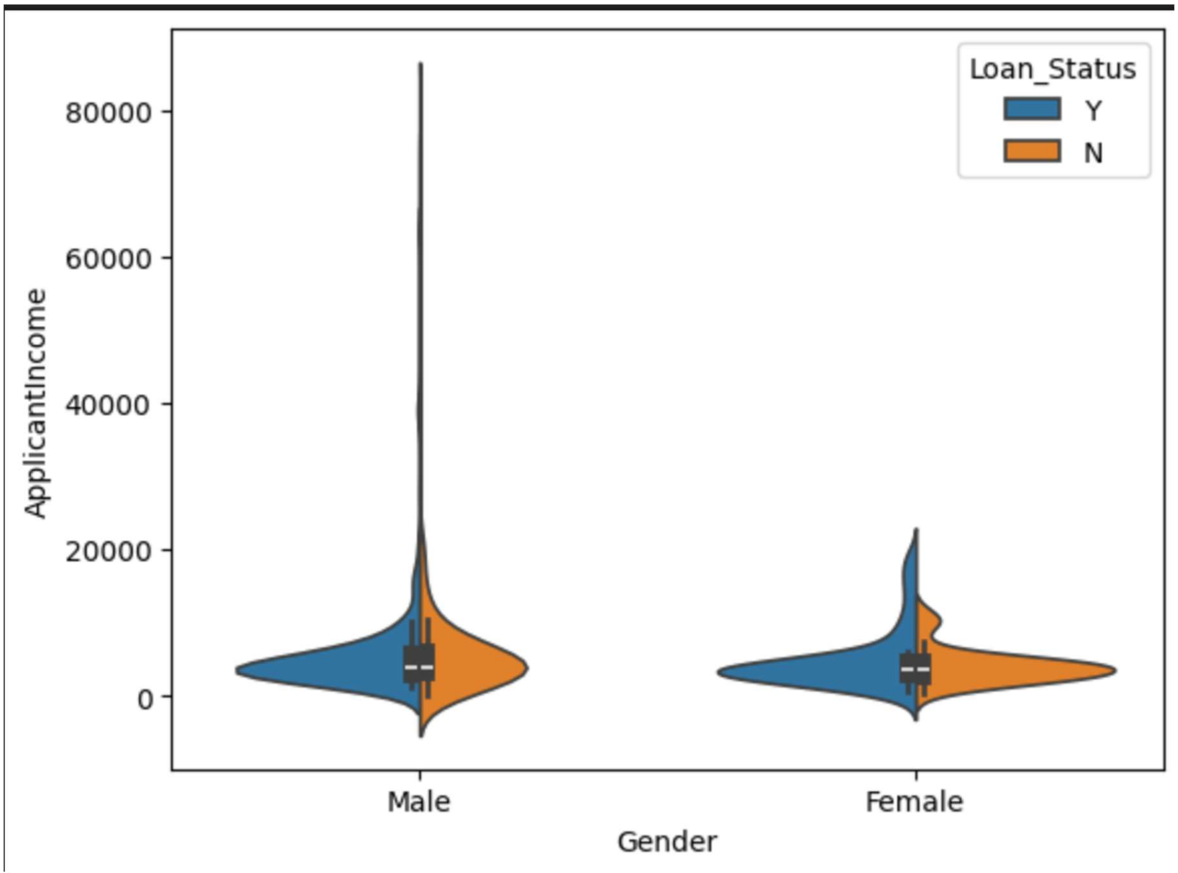
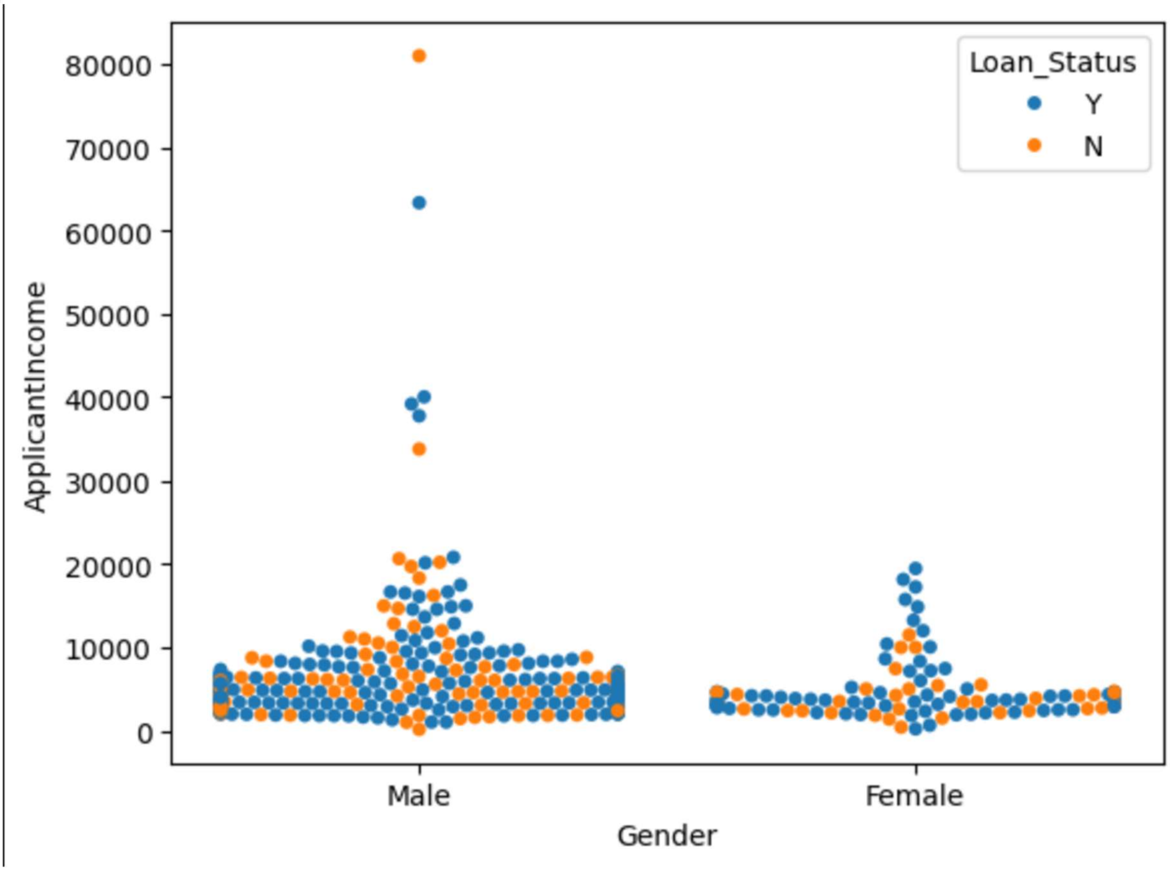
	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_Hist
count	614	601	611	599	614	582	614.000000	614.000000	592.000000	600.00000	564.000
unique	614	2	2	4	2	2	NaN	NaN	NaN	NaN	N
top	LP001002	Male	Yes	0	Graduate	No	NaN	NaN	NaN	NaN	N
freq	1	489	398	345	480	500	NaN	NaN	NaN	NaN	N
mean	NaN	NaN	NaN	NaN	NaN	NaN	5403.459283	1621.245798	146.412162	342.00000	0.842
std	NaN	NaN	NaN	NaN	NaN	NaN	6109.041673	2926.248369	85.587325	65.12041	0.364
min	NaN	NaN	NaN	NaN	NaN	NaN	150.000000	0.000000	9.000000	12.00000	0.000
25%	NaN	NaN	NaN	NaN	NaN	NaN	2877.500000	0.000000	100.000000	360.00000	1.000
50%	NaN	NaN	NaN	NaN	NaN	NaN	3812.500000	1188.500000	128.000000	360.00000	1.000
75%	NaN	NaN	NaN	NaN	NaN	NaN	5795.000000	2297.250000	168.000000	360.00000	1.000
max	NaN	NaN	NaN	NaN	NaN	NaN	81000.000000	41667.000000	700.000000	480.00000	1.000

<table>
<tr>
<td>Univariate Analysis</td>
<td>


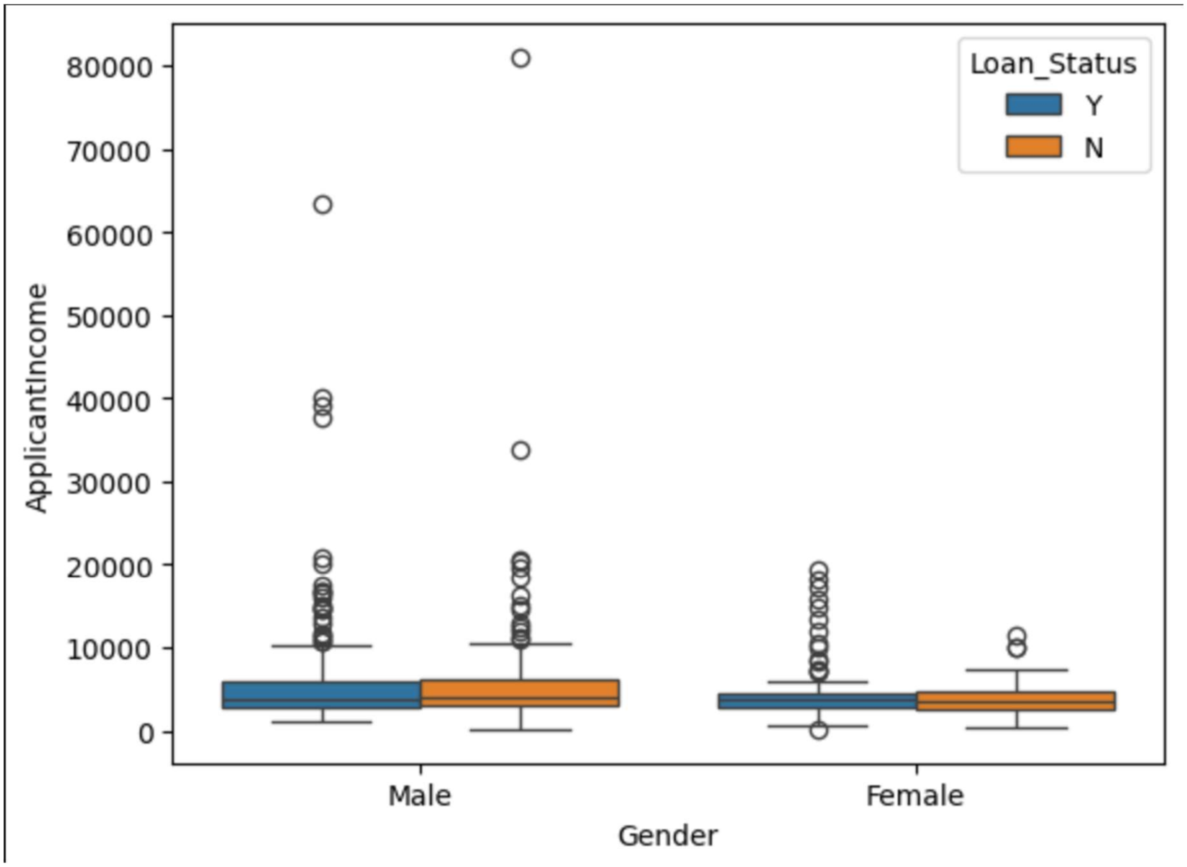
</td>
</tr>
<tr>
<td>Bivariate Analysis</td>
<td>




</td>
</tr>
</table>

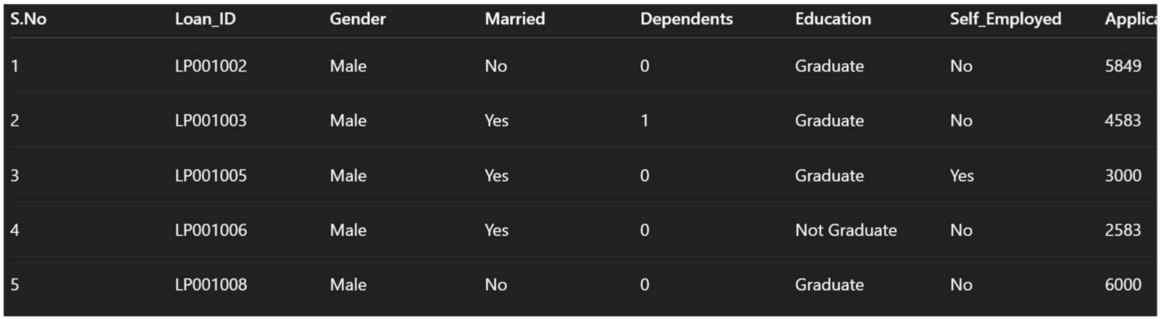
Multivariate Analysis	 

Outliers and Anomalies



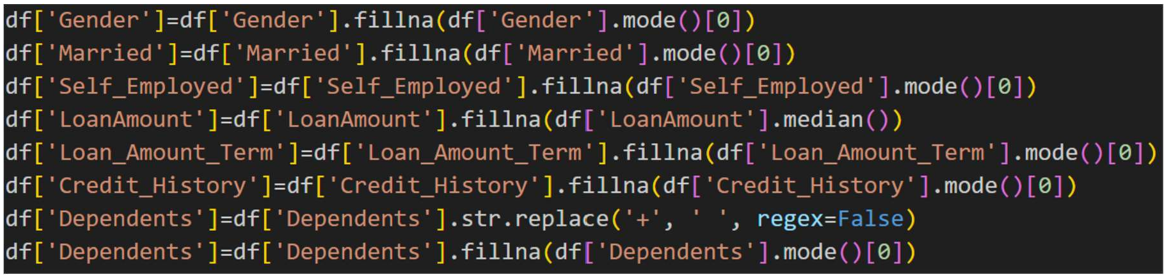
Data Preprocessing Code Screenshots

Loading Data



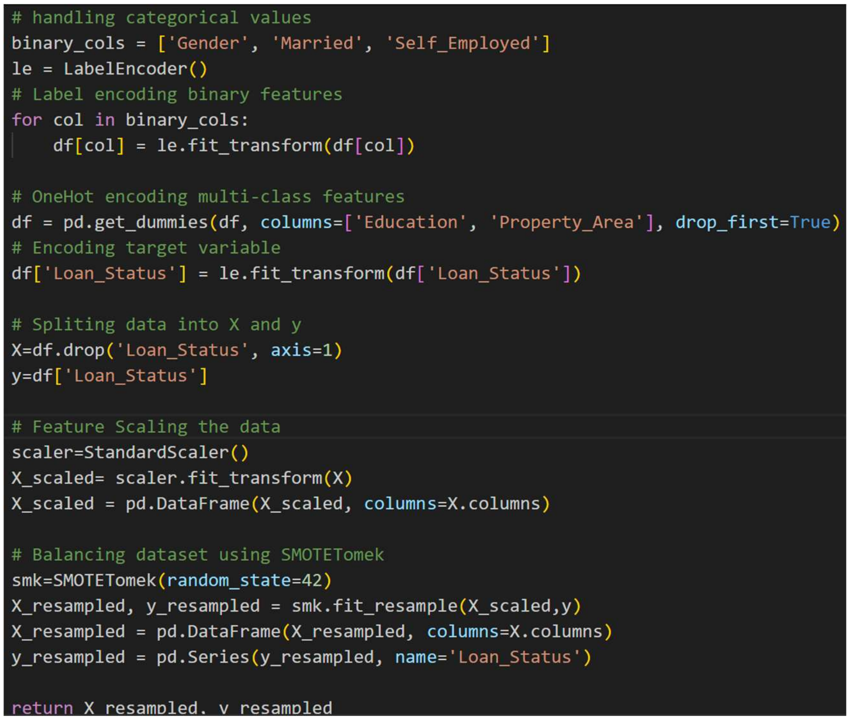
S.No	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	Applica
1	LP001002	Male	No	0	Graduate	No	5849
2	LP001003	Male	Yes	1	Graduate	No	4583
3	LP001005	Male	Yes	0	Graduate	Yes	3000
4	LP001006	Male	Yes	0	Not Graduate	No	2583
5	LP001008	Male	No	0	Graduate	No	6000

Handling Missing Data



```
df['Gender']=df['Gender'].fillna(df['Gender'].mode()[0])
df['Married']=df['Married'].fillna(df['Married'].mode()[0])
df['Self_Employed']=df['Self_Employed'].fillna(df['Self_Employed'].mode()[0])
df['LoanAmount']=df['LoanAmount'].fillna(df['LoanAmount'].median())
df['Loan_Amount_Term']=df['Loan_Amount_Term'].fillna(df['Loan_Amount_Term'].mode()[0])
df['Credit_History']=df['Credit_History'].fillna(df['Credit_History'].mode()[0])
df['Dependents']=df['Dependents'].str.replace('+', ' ', regex=False)
df['Dependents']=df['Dependents'].fillna(df['Dependents'].mode()[0])
```

<table>
<tr>
<td>Data Transformation</td>
<td>



```
# handling categorical values
binary_cols = ['Gender', 'Married', 'Self_Employed']
le = LabelEncoder()
# Label encoding binary features
for col in binary_cols:
    df[col] = le.fit_transform(df[col])

# OneHot encoding multi-class features
df = pd.get_dummies(df, columns=['Education', 'Property_Area'], drop_first=True)
# Encoding target variable
df['Loan_Status'] = le.fit_transform(df['Loan_Status'])

# Spliting data into X and y
X=df.drop('Loan_Status', axis=1)
y=df['Loan_Status']

# Feature Scaling the data
scaler=StandardScaler()
X_scaled= scaler.fit_transform(X)
X_scaled = pd.DataFrame(X_scaled, columns=X.columns)

# Balancing dataset using SMOTETomek
smk=SMOTETomek(random_state=42)
X_resampled, y_resampled = smk.fit_resample(X_scaled,y)
X_resampled = pd.DataFrame(X_resampled, columns=X.columns)
y_resampled = pd.Series(y_resampled, name='Loan_Status')

return X_resampled, y_resampled
```

</td>
</tr>
<tr>
<td>Feature Engineering</td>
<td>Attached the codes in final submission.</td>
</tr>
<tr>
<td>Save Processed Data</td>
<td>-</td>
</tr>
</table>