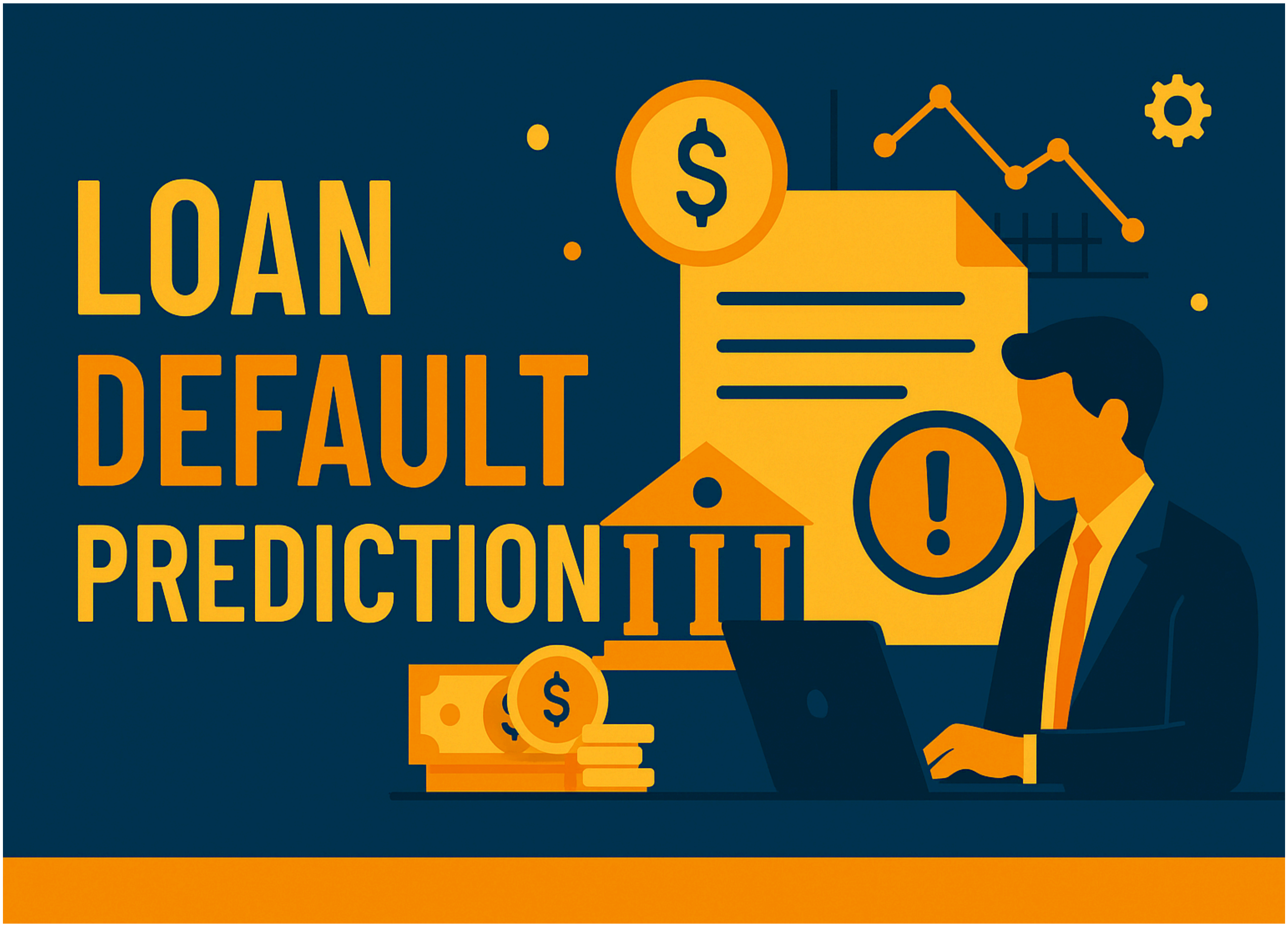


KAPIL CHHETRI
09/04/2024

Tools Used

- **Pandas :-** Data Manipulation, Exploratory Data Analysis
- **Matplotlib & Seaborn :-** Data Visualizations
- **Scikit-Learn:-** train_test_split, StandardScaler, LogisticRegression, Confusion_Matrix
- MS PowerBi: - Visualization & Dashboard

Data Source

DataLab: - loan_data.csv(https://www.datacamp.com/datalab/datasets/dataset-python-loans)

Objectives

- **How many customers defaulted on their loan?**
- **Why do customers default on their loans?**
- **How many customers going to default their loan in future?**
- **Analyse and visualize the insights.**

Stages of Analysis

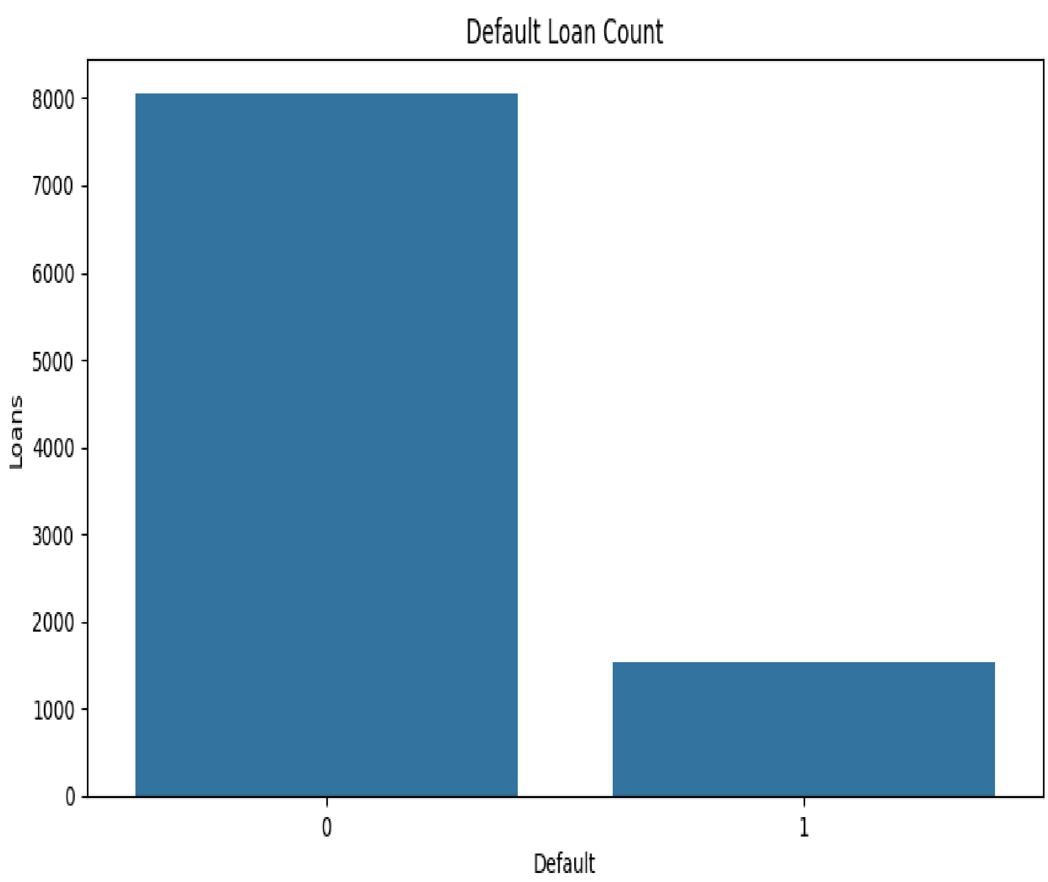
Step 1:- Data Observation

- df.info()
- df.describe()
- df.isnull().sum()
- df['not.fully.paid'].value_counts()

Step 2:- Exploratory Data Analysis

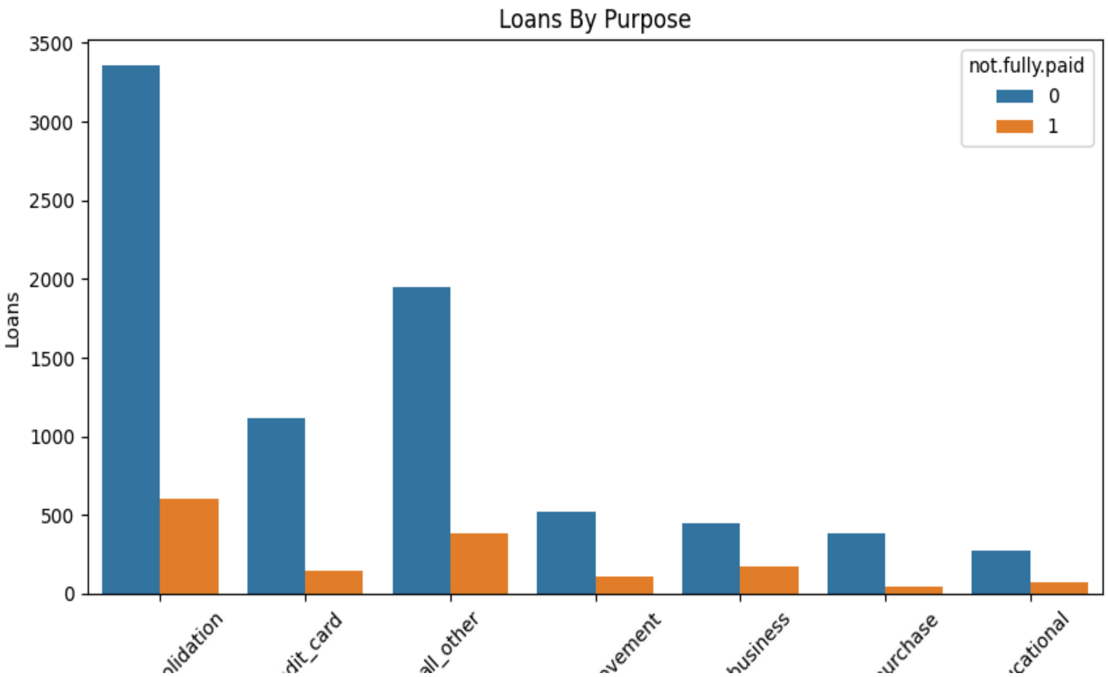
- **Loan Default Distribution: -**

```
plt.figure(figsize = (10, 5))
sns.countplot(x = 'not.fully.paid', data = df)
plt.title('Default Loan Count')
plt.xlabel('Default')
plt.ylabel('Loans')
plt.savefig("loan_default_distribution.png")
plt.show()
```



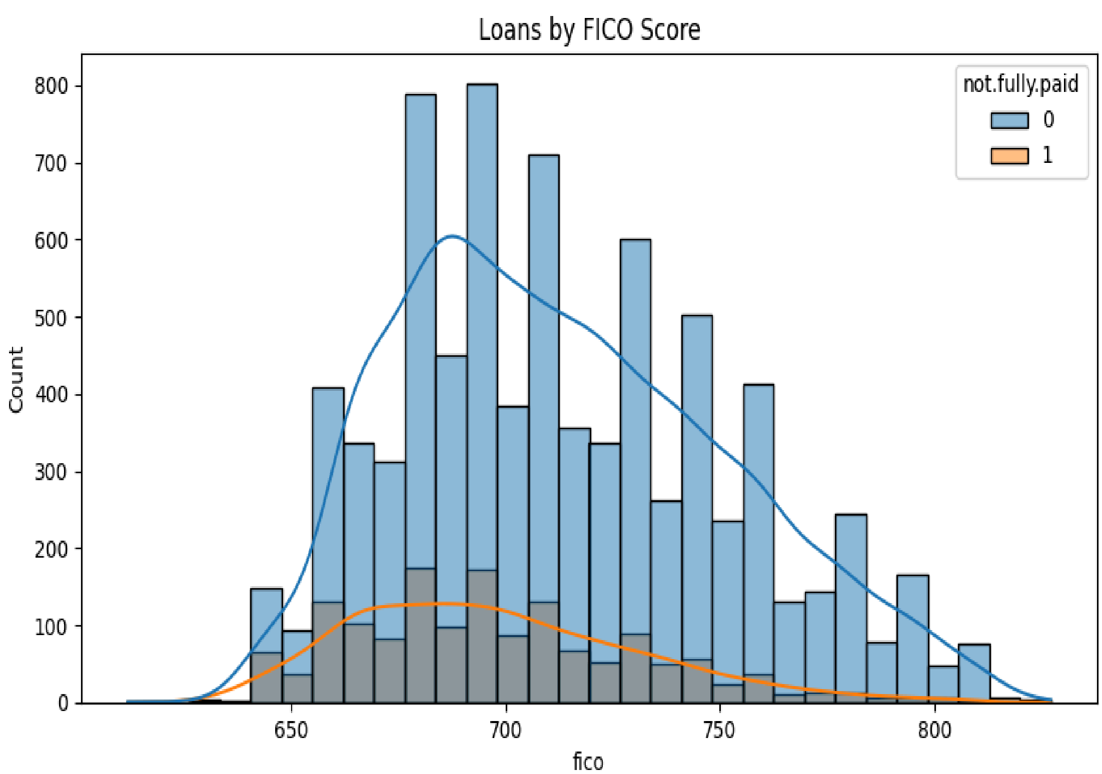
- **Loan by Purpose:-**

```
plt.figure(figsize = (10, 5))
sns.countplot(x ='purpose', hue= 'not.fully.paid', data = df)
plt.title('Loans By Purpose')
plt.xlabel('Purpose')
plt.xticks(rotation = 45)
plt.ylabel('Loans')
plt.savefig("loan_default_by_purposse.png")
plt.show()
```



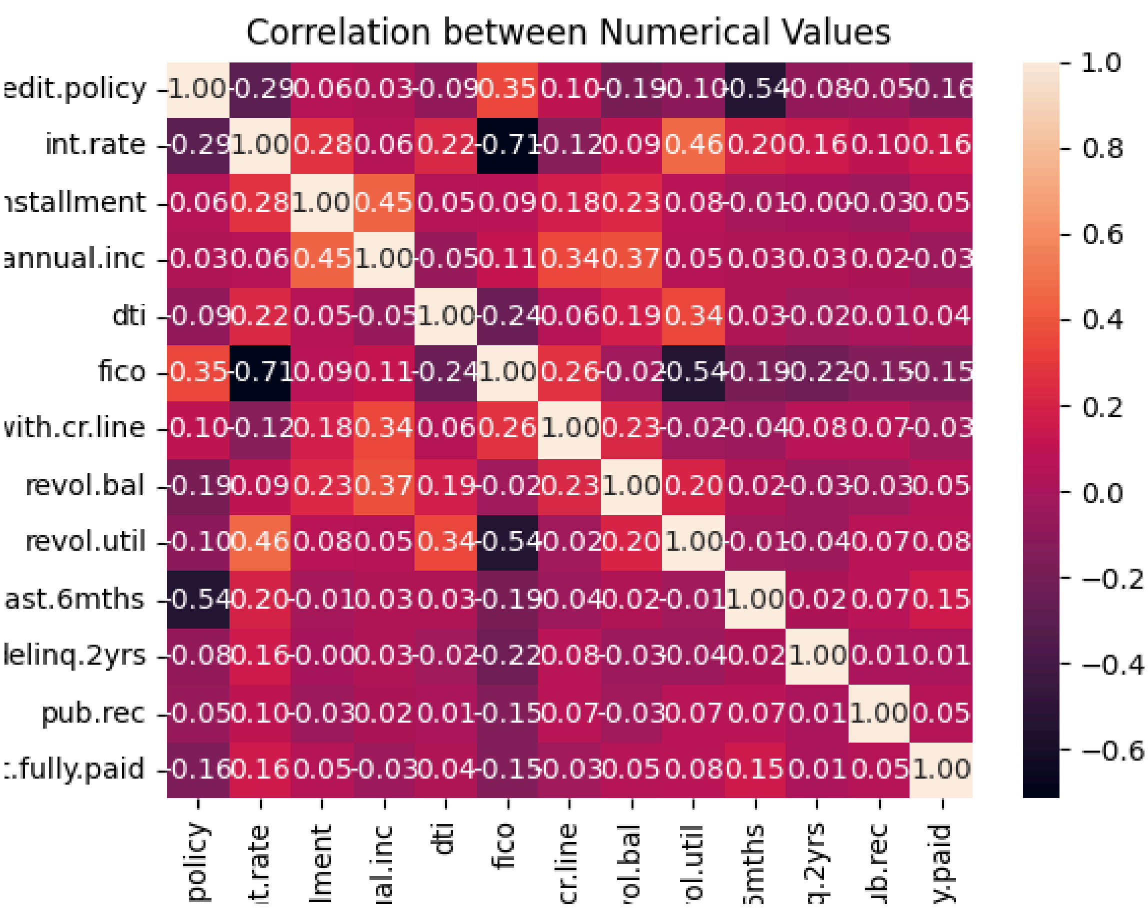
- **Loans By FICO Score: -**

```
plt.figure(figsize = (10, 5))

sns.histplot(x = 'fico', hue = 'not.fully.paid', data = df, bins = 30, kde = True)

plt.title('Loans by FICO Score')

plt.savefig("loan_default_by_FICO.png")

plt.xlabel('FICO Score')

plt.ylabel('Loans')

plt.show()
```



- **Correlation Between Features: -**

```
sns.heatmap(df_num.corr(), annot = True, cmap = 'rocket', fmt ='.2f')

plt.title('Correlation between Numerical Values')

plt.savefig("Datset_Correlation.png")

plt.show()
```



Step 3:- Feature Engineering

- **Encoding: -**

  df_encoded = pd.get_dummies(df, columns = ['purpose'], drop_first = True)

- **Feature Selection: -**

  x = df_encoded.drop(['not.fully.paid'], axis = 1)

  y = df['not.fully.paid']

- **train_test_split:-**

  x_train, x_test, y_train, y_test = train_test_split(x_scaled, y, test_size = 0.2, random_state = 42)

Step 4:- ML Modelling

- **LogisticRegression: -**

  model = LogisticRegression(max_iter = 1000, solver = 'liblinear', class_weight = 'balanced')

  model.fit(x_train, y_train)

  y_pred = model.predict(x_test)

  results = pd.DataFrame({

  'Actual': y_test.values,

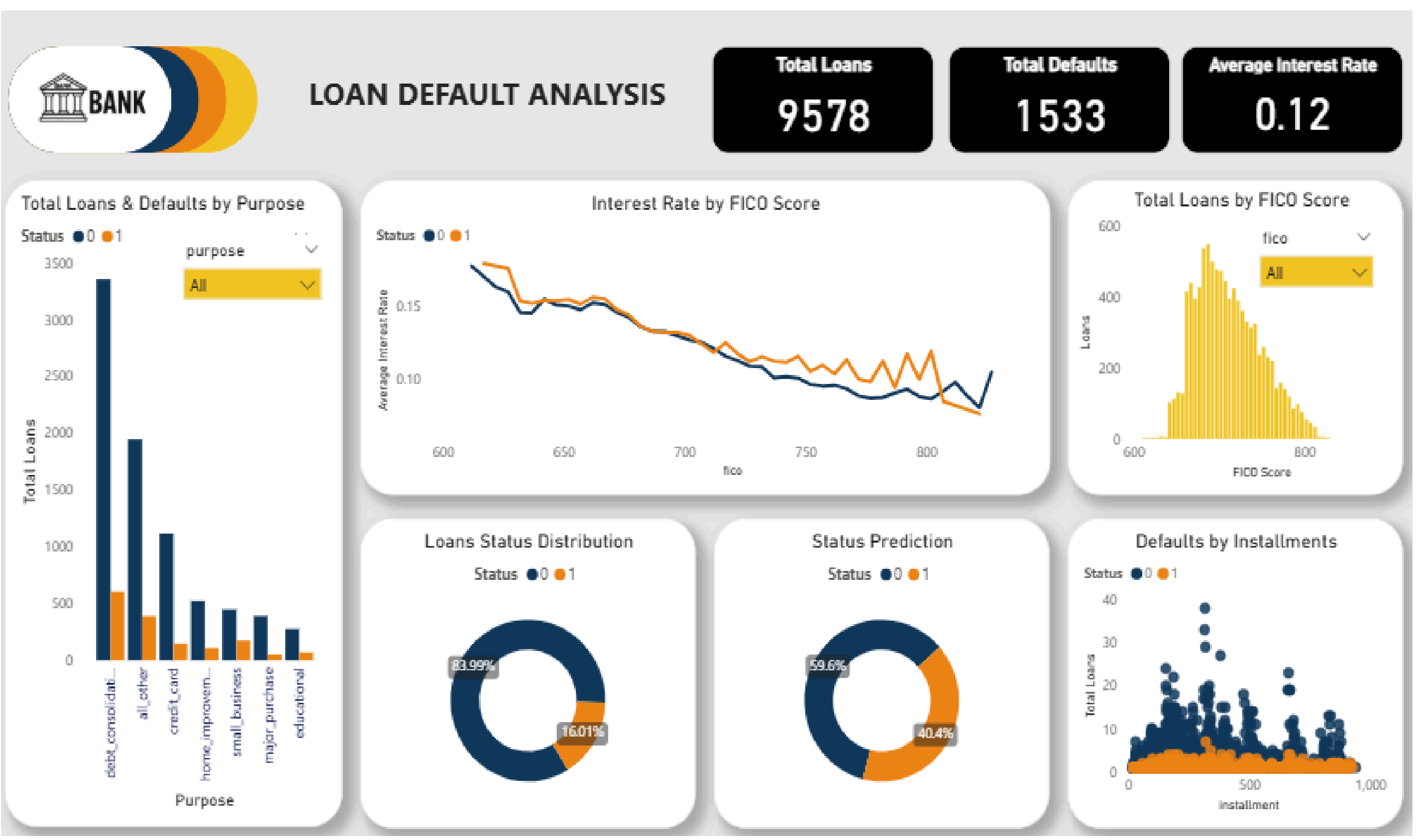
  'Predicted': y_pred

  })

Step 5:- Exporting Analysis

```
results = pd.DataFrame({
    'Actual': y_test.values,
    'Predicted': y_pred,
    'FICO': x_test[:, list(x.columns).index('fico')],  # example
    'Interest Rate': x_test[:, list(x.columns).index('int.rate')],
})

results.to_csv("loan_predictions_results.csv", index=False)
```

Step 6:- MS PowerBI Dashboard:-



Key Insights

- The majority of individuals are taking out loans for debt consolidation, with the total exceeding 3,000 loans.
- The loan purpose categorized as "major purchase" has the lowest default rate relative to the total number of loans.
- A higher FICO score is associated with fewer loan defaults.
- Currently, 16.01% of loans have defaulted, and this rate is projected to increase to 40.4% in the future.