University of Lancashire

Case Study – Loan Prediction

Karen O'Shea

Where opportunity creates success

Case Study Example

Loan Predication/Approval Dataset – provide guidance for summative assessment

- Example Data Science project ‘journey’, including:
  - Examining datasets – training and testing
  - Asking questions and proposing hypotheses
  - Visualise data
  - Identifying outliers
  - Consider algorithm design

- Algorithms – Semester 2

Problem Statement – Loan Prediction

Finance Associates deal in home loans. Customers apply for a loan which is evaluated and validated for eligibility. The company wants to automate the process in 'real-time', which will include details such as: gender, marital status, education, number of dependants, income, credit history and loan amount. Which customer features need identifying to target those that are eligible?

- Which customer features need identifying to target those that are eligible?

Loan Prediction Problem

University of Lancashire

Classification Problem

- Binary (yes/no; win/loss...) or,
- Multiclass (classifying groups ie. breads of animals; categories of movies...)
- What is your hypothesis; prediction or research question?

Read/Examine Dataset

- Train and Test datasets (taken from kaggle.com)

  train=pd.read_csv(“train.csv”)

- Examine the structure of the datasets – how many variables and type of data? Target variable: Loan_Status

Understand/Describe Dataset

Loan ID; Gender; Married; Dependents; Education; Self-Employed; Applicant Income; Co Applicant Income; Loan Amount; Loan Amount Term; Credit History; Property Area; Loan Status

Mixture of categorical, ordinal and numerical fields

Variable	Description
Loan_ID	Unique Loan ID
Gender	Male/ Female
Married	Applicant married (Y/N)
Dependents	Number of dependents
Education	Applicant Education (Graduate/Under Graduate)
Self_Employed	Self employed (Y/N)
ApplicantIncome	Applicant income
CoapplicantIncome	Coapplicant income
LoanAmount	Loan amount in thousands
Loan_Amount_Term	Term of loan in months
Credit_History	Credit history meets guidelines
Property_Area	Urban/ Semi Urban/ Rural
Loan_Status	Loan approved (Y/N)

Recap: Import Libraries

Import libraries:

- pandas for dataframes
- numpy for calculations
- seaborn for visualisations
- matplotlib for plotting graphs

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
```

```
train=pd.read_csv("train.csv")
```

https://pandas.pydata.org/docs/getting_started/intro_tutorials/01_table_oriented.html

Univariant Analysis – Target Variable

- Target Variable – Loan Status
- Count number of approved loans
- Normalize outcomes ie. proportions (0-1 range)

Task:

- Import libraries
- Read in ‘Train’ dataset
- Examine dataset and
- Count ‘Loan Status’

```
train['Loan_Status'].value_counts()
```

```
Y    422
N    192
Name: Loan_Status, dtype: int64
```

```
train['Loan_Status'].value_counts(normalize=True)
```

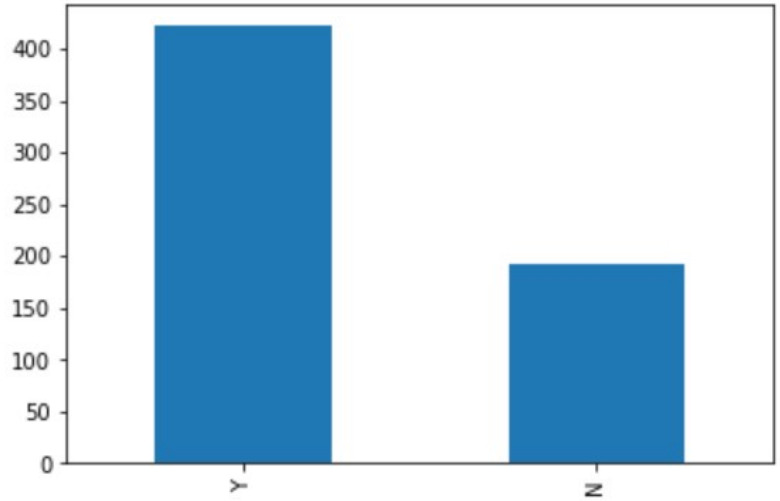
```
Y    0.687296
N    0.312704
Name: Loan_Status, dtype: float64
```

Univariant Analysis – Visualise Data

University of Lancashire

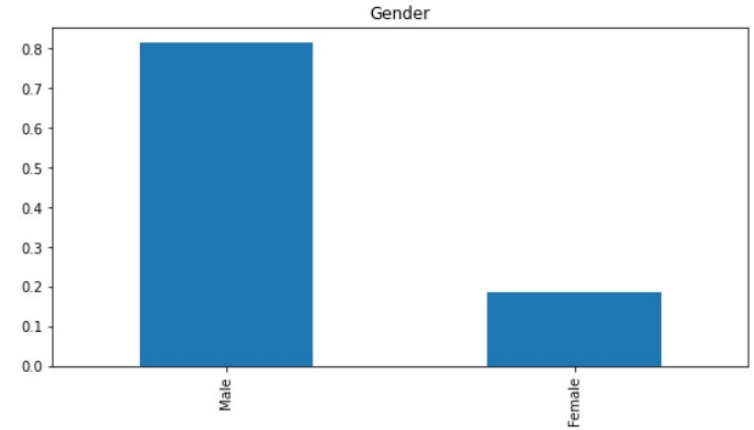
```
train['Loan_Status'].value_counts().plot.bar()
```

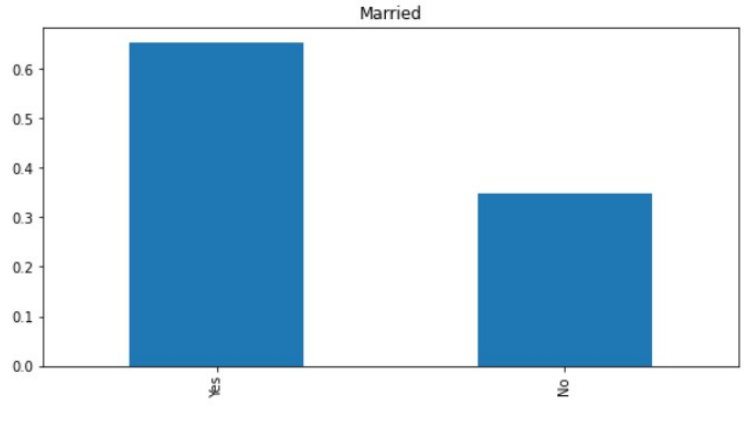
```
]: <AxesSubplot:>
```



Sub Plots – Categories: Gender and Married

```
plt.subplot(221)
train['Gender'].value_counts(normalize=True).plot.bar(figsize=(20, 10), title='Gender')
plt.subplot(222)
train['Married'].value_counts(normalize=True).plot.bar(title='Married')
plt.show()
```





Further Categorical Features Analysis

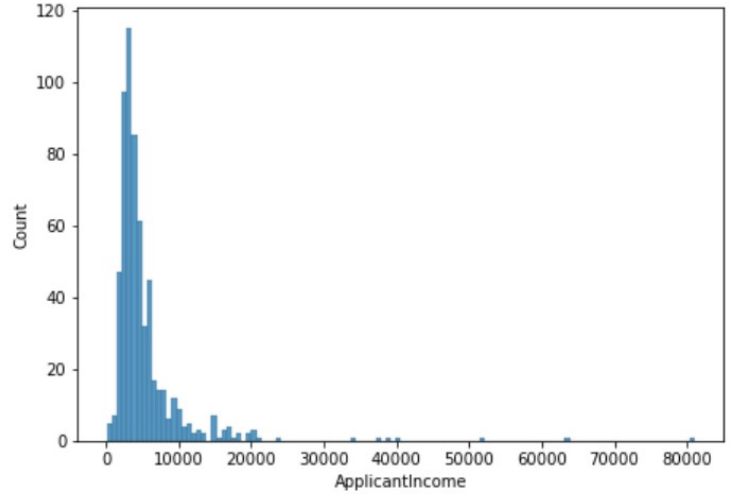
University of Lancashire

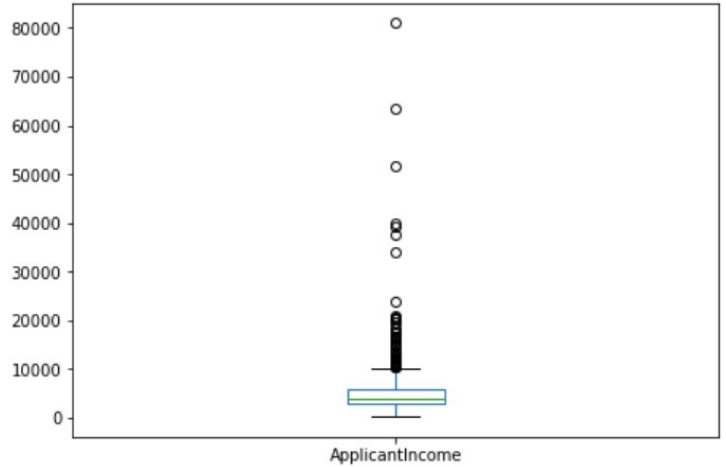
Task: visualise other categorical features vs Loan Status:

- Gender, Married, Self-Employed, Credit History, Education, Dependents
- What can be inferred?

Visualise Numerical Features

```
plt.figure(1)
plt.subplot(121)
sns.histplot(train['ApplicantIncome']);
plt.subplot(122)
train['ApplicantIncome'].plot.box(figsize=(16, 5))
plt.show()
```



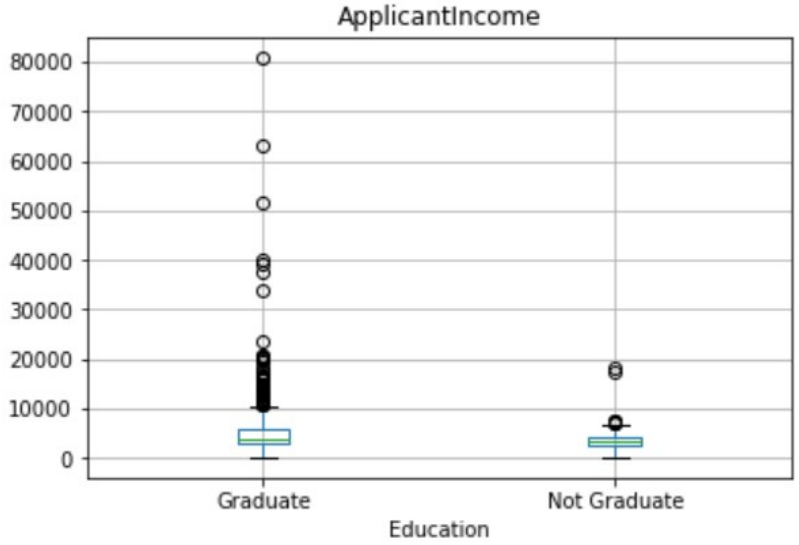


Numerical Features: Segregate by Education

University of Lancashire

```
train.boxplot(column='ApplicantIncome', by='Education')
plt.suptitle("")
```

```
Text(0.5, 0.98, '')
```



Further Numerical Features Analysis

University of Lancashire

- **Task**: visualise numerical features:
  - Co applicant's income and loan amount

- Do we see a normal distribution?
- Any outliers?
- Can we make any assumptions?

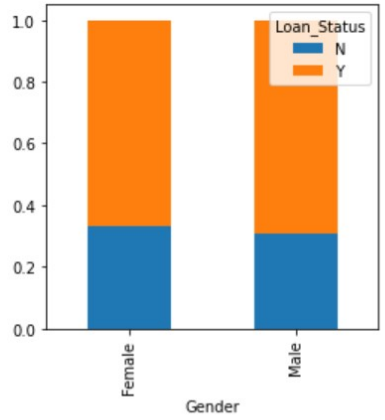
Bivariant Analysis – Categorical Features

- Relationship between target variable (Loan Status) and Gender

- **Task**: visualise other categorica variables including:
- Married
- Dependents
- Education
- Self Employed

```
Gender=pd.crosstab(train['Gender'],train['Loan_Status'])
Gender.div(Gender.sum(1).astype(float), axis=0).plot(kind="bar", stacked=True, figsize=(4, 4))
```

```
]: <AxesSubplot:xlabel='Gender'>
```

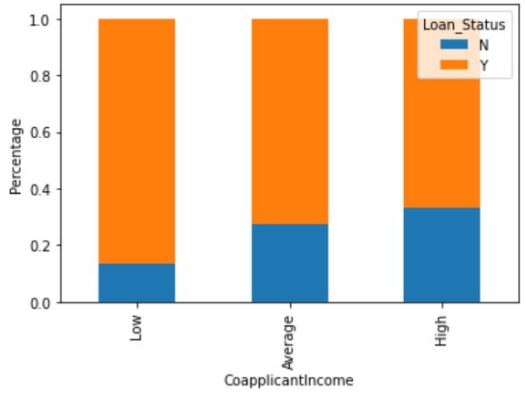


Bivariant Analysis – Numerical Features

- Relationship between target variable (Loan Status) and Co-applicant Incomes
- Does this support our hypothesis?
- Is loan approval dependent on a co-applicant

```
bins=[0, 1000, 3000, 42000]
group=['Low', 'Average', 'High']
train['Coapplicant_Income_bin']=pd.cut(train['CoapplicantIncome'], bins,labels=group)

Coapplicant_Income_bin=pd.crosstab(train['Coapplicant_Income_bin'],train['Loan_Status'])
Coapplicant_Income_bin.div(Coapplicant_Income_bin.sum(1).astype(float), axis=0).plot(kind="bar", stacked=True)
plt.xlabel('CoapplicantIncome')
P=plt.ylabel('Percentage')
```



Combining Variables (Combined Income) for Analysis

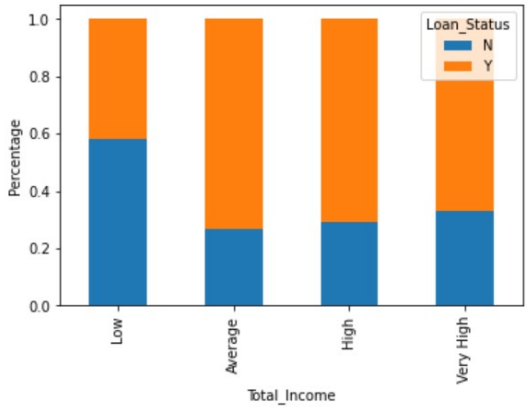
University of Lancashire

- Relationship between target variable (Loan Status) and Combined Total Income (Applicant and Co-applicant)

- Does this provide further insight and prove our hypothesis?

```
train['Total_Income']=train['ApplicantIncome']+train['CoapplicantIncome']
bins=[0,2500,4000,6000,81000]
group=['Low', 'Average', 'High', 'Very High']
train['Total_Income_bin']=pd.cut(train['Total_Income'],bins,labels=group)

Total_Income_bin=pd.crosstab(train['Total_Income_bin'],train['Loan_Status'])
Total_Income_bin.div(Total_Income_bin.sum(1).astype(float), axis=0).plot(kind="bar", stacked=True)
plt.xlabel('Total_Income')
P=plt.ylabel('Percentage')
```



Further Bivariate Numerical Analysis

- **Task**: Visualise Loan Amount
  - What proportions of loans are approved?

- In summary: can you draw any correlations between the categorical and numerical variables analysed?

Assignment 1 Review: Consider....