University of Lancashire

Case Study (2) – Loan Prediction

Karen O'Shea

Where opportunity creates success

Learning Objectives

- Discuss problem statement by proposing hypotheses
- Create data visualisations using selected features of a dataset
- Evaluate data visualisations reflecting on hypotheses and problem statement

Case Study Example

Loan Predication/Approval Dataset – provide guidance for summative assessment

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

- Algorithms – Semester 2

Questions/Hypothesis

University of Lancashire

What can effect loan approval:

- All customers over the age of 50 will be accepted for a loan
- Loan acceptance requires higher levels of incomes

Task: Refer back to problem statement. Any other examples...

Recap: Understand/Describe Dataset

University of Lancashire

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)

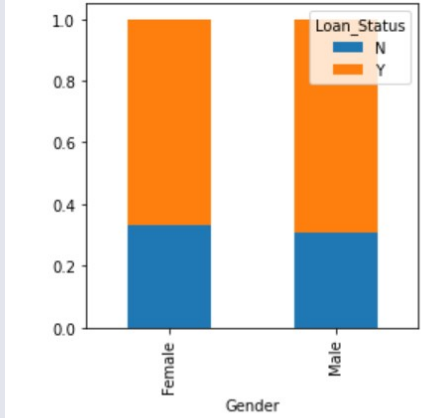
Recap: Bivariant Analysis – Categorical Features

- 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'>
```

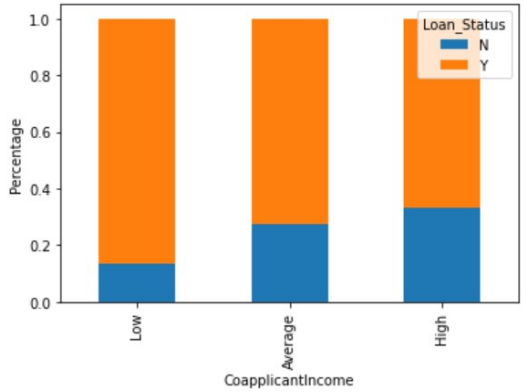


Recap: Bivariant Analysis – Numerical Features

- 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')
```



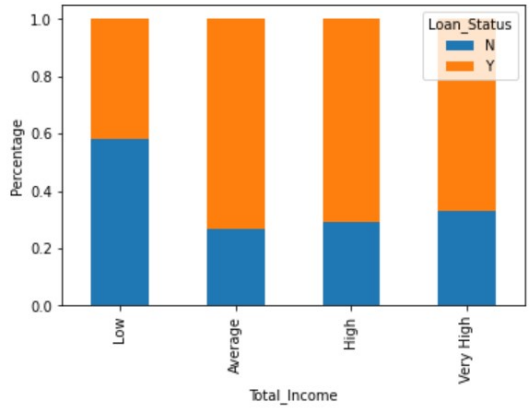
Recap: Combining Variables (Combined Income) for Analysis

- 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')
```



Missing/Outlier Data?

- Impact of missing data and outliers

- Have you identified any missing data?

Feature-wise: count of missing data

- There are missing values in all features
- Consider numerical and categorical features separately
- Imputation using mean, median and mode

```
train.isnull().sum()
```

```
]: Loan_ID              0
   Gender              13
   Married              3
   Dependents          15
   Education            0
   Self_Employed       32
   ApplicantIncome      0
   CoapplicantIncome    0
   LoanAmount          22
   Loan_Amount_Term    14
   Credit_History      50
   Property_Area        0
   Loan_Status          0
   dtype: int64
```

Filling Missing Values (1)

- There are some other categories with missing values...
- Could generalise here using the mode

Task:

- Consider all categories with missing data and amend?
- Make appropriate judgements here...

```
train['Gender'].fillna(train['Gender'].mode()[0], inplace=True)
```

Filing Missing Values (2)

- Loan Amount has missing values – could we use mean here?
- Are there outliers which could impact the value of the mean?
- What about other statistical measures such as median?

```
train['LoanAmount'].fillna(train['LoanAmount'].median(), inplace=True)
```

- Carry out a check to see if there are any other missing values

Loan Amount – Normal Distribution

- Evidence of outliers? Use a histogram to view distribution. Is the distribution symmetric or skewed?

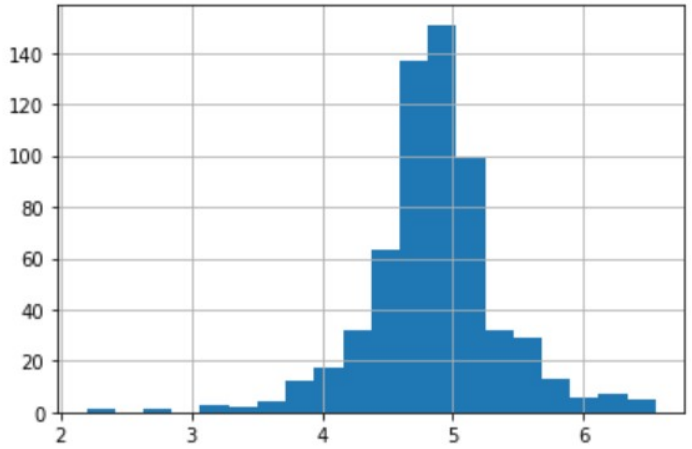
- Attempt ‘Log Transformation’ to produce a ‘more’ normal distribution. Does not affect smaller values but does reduce larger values.

Log Transformation

University of Lancashire

```
train['LoanAmount_log'] = np.log(train['LoanAmount'])
train['LoanAmount_log'].hist(bins=20)
```

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

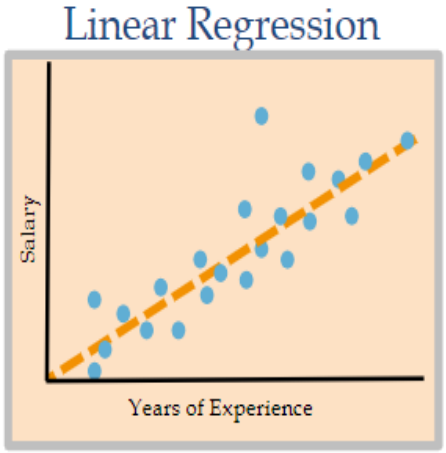


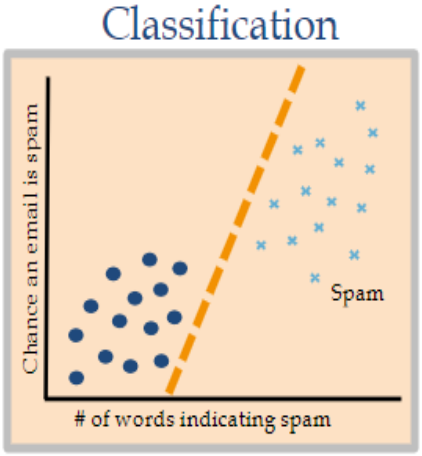
Semester 2: Next Steps…

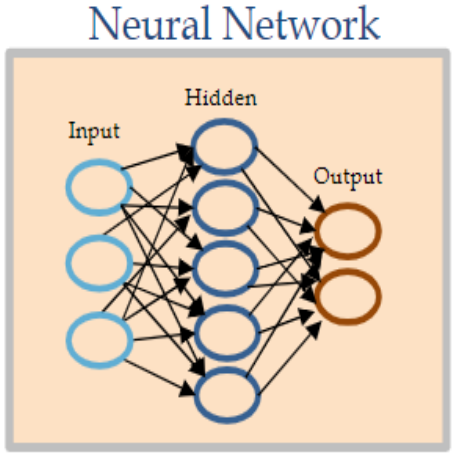
- Build a Data Science Model…examples include: Linear/Logistic Regression

- Build and make predictions using ‘Test’ dataset

Consider....


Linear Regression
Salary
Years of Experience

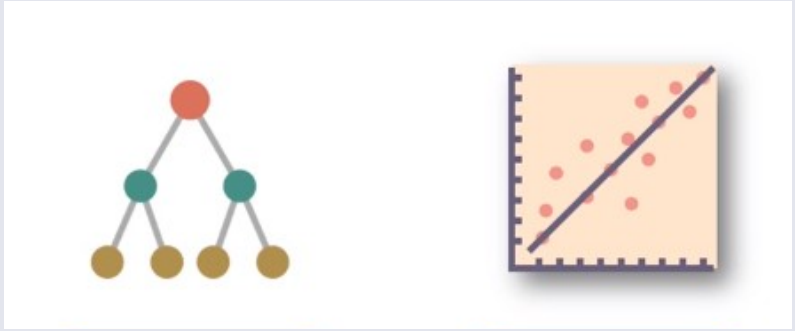

Classification
Chance an email is spam
Spam
# of words indicating spam


Neural Network
Hidden
Input
Output

Data Modelling Algorithms

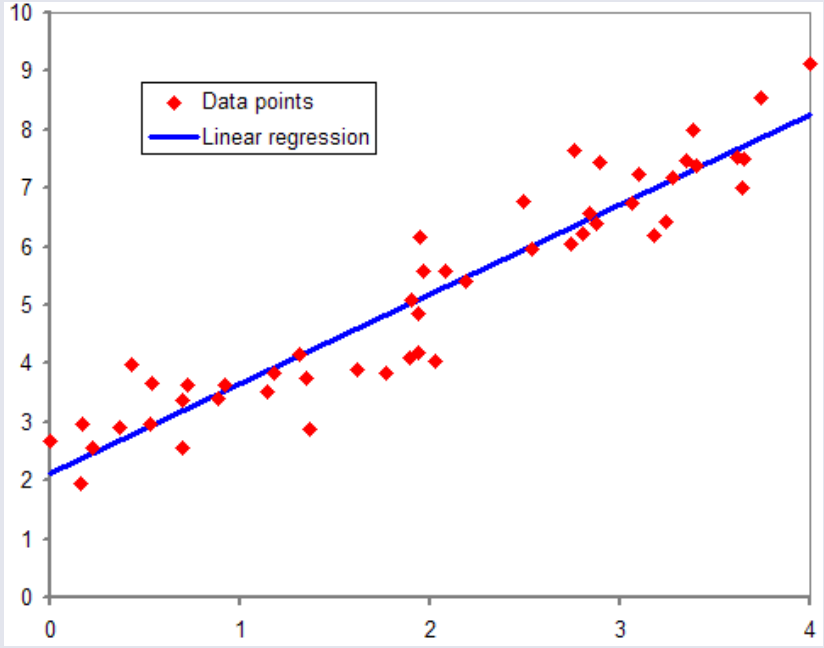
Question:

- Linear Regression
- Logistic Regression (Classification)
- Decision Tree (Random Forests)
- Unsupervised learning (Clustering)



Discuss: What do we know so far from our research?

Linear Regression – Supervised Learning



Equation of a line:

y = Mx + C

M = gradient
C = Intercept

Line of ‘best’ fit

Assignment 2: Next Steps…

Keep asking the questions:

- Does my data make sense?
- Is the data consistent?
- What do you make of the data's distribution? Does it change over time…is this to be expected?
- Use visualisations here to help – do you need to normalise data first?
- Is the data complete…missing values or anomalies?
- Do you understand the features…any data transformations required (data types)?
- Balanced or unbalanced data? Require a 50/50 split. Consider 'undersampling' or 'oversampling'.
- Any additional data that might be beneficial?

Assignment 2: Future steps…

Consider:

- Metric for evaluation – what is meant by a ‘good’ model?

- Splitting data – training and testing sets (import train_test_split).

Learning Objectives

- Discuss problem statement by proposing hypotheses
- Create data visualisations using selected features of a dataset
- Evaluate data visualisations reflecting on hypotheses and problem statement