alary Prediction

Income prediction - using decision tree

Grace Palma

Problem

- Most valuable question in Market research:

Personal income

- Potential spending habits
- Discretionary income
- Audience segregation
- Product or campaign targeting

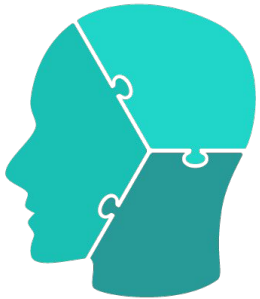
Background

- ABS census: Mean yearly personal income
  - $80,000 +
- Aim
  - To predict if a person earns above or below the average salary for Australians

Data

	Gender	State	Age	Employment	Partnership	Children	No_of_children	Home_status	Qualified	Income	Salary	Work_type	BDM	Industry
0	1	2	22	1	0	0	0	NaN	1	102499.5	Above 80	2	1	16
1	0	0	0	0	0	0	0	NaN	1	0.0	NaN	0	0	0
2	2	1	57	1	1	1	0	Mortgage	1	0.0	NaN	1	1	22
3	1	1	42	1	0	0	0	NaN	1	352499.5	Above 80	1	1	23
4	1	1	42	1	0	0	0	Renting	0	0.0	NaN	2	1	24
5	0	0	0	0	0	0	0	NaN	1	0.0	NaN	0	0	0
6	0	0	0	0	0	0	0	NaN	1	0.0	NaN	0	0	0
7	2	1	37	0	0	0	0	NaN	0	0.0	NaN	0	0	0
8	1	1	67	0	0	0	0	NaN	0	0.0	NaN	0	0	0
9	1	2	57	1	0	0	0	Mortgage	0	92499.5	Above 80	2	1	19

- Survey data
- Online panel
- Contains information on an individual's relationship status, work related information, etc.



Tools

- from sklearn import tree
- import pandas as pd
- import numpy as np
- import math

Visualisation:

- import matplotlib.pyplot as plt
- import seaborn as sns
- %matplotlib inline

Change categorical to numerical:

- for c in features_categorical:

  df[c] = pd.Categorical(df[c]).codes

Model classifier:

- from sklearn.tree import DecisionTreeClassifier

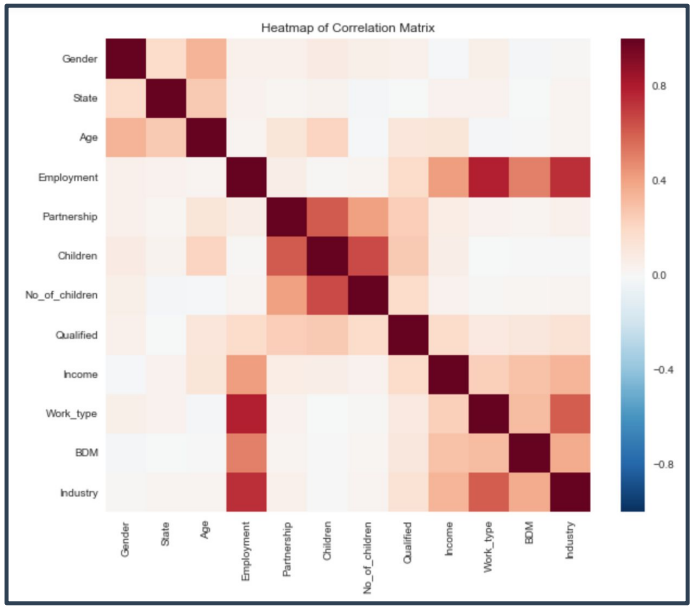
Accuracy test:

- from sklearn.metrics import roc_auc_score

Feature selection

- Correlation matrix
- Heatmap of correlation matrix

	Income
Employment	0.418967
Industry	0.342959
BDM	0.289510
Work_type	0.238678
Qualified	0.191781
Age	0.130094
Partnership	0.077290
Children	0.067124
State	0.045941
No_of_children	0.039495
Gender	-0.008289

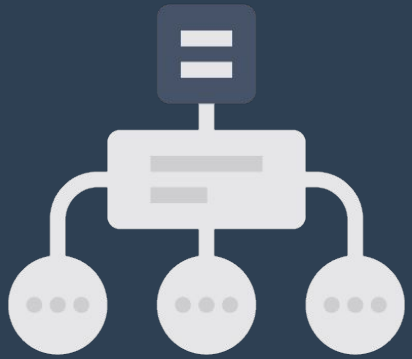


Pre-processing

- Define relevant variables or features
- Define categorical variables
- Change categorical variables to numeric
- Define features
- Define target

	Employment	Industry	BDM	Work_type	Age	Qualified	Partnership	Salary	Home_status	Gender
0	1	16	1	2	22	1	0	1	0	1
1	0	0	0	0	0	1	0	0	0	0
2	1	22	1	1	57	1	1	0	2	2
3	1	23	1	1	42	1	0	1	0	1
4	1	24	1	2	42	0	0	0	4	1
5	0	0	0	0	0	1	0	0	0	0
6	0	0	0	0	0	1	0	0	0	0
7	0	0	0	0	37	0	0	0	0	2
8	0	0	0	0	67	0	0	0	0	1
9	1	19	1	2	57	0	0	1	2	1

Analysis

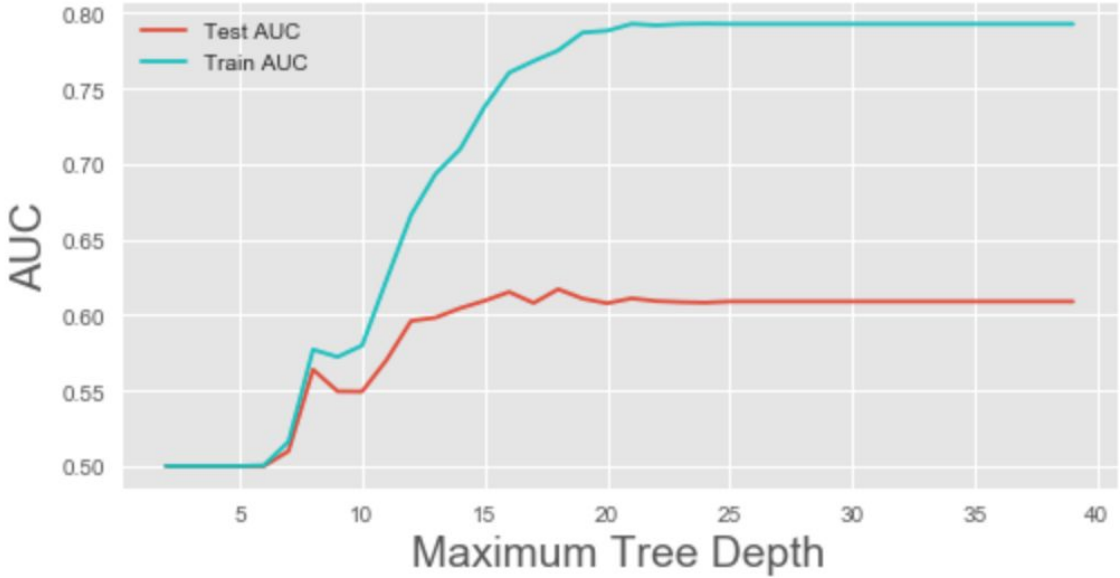


- Split and train the data set
  - 80% for training
  - 20% for testing
- DecisionTreeClassifier from scikit learn package to fit the decision tree

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_split=1e-07, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            presort=False, random_state=1, splitter='best')
```

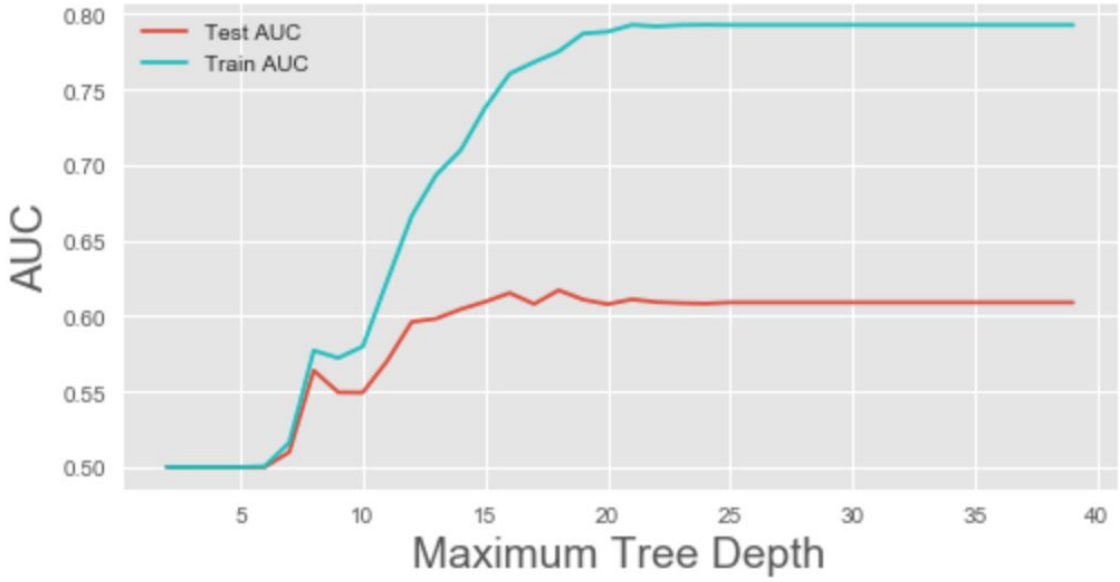
Accuracy test

- Area under the curve
  - Test: 0.608960029048
  - Training: 0.792767534575
- Tree depth restrictions



Tree depth restrictions

- Depth restricted to 18
- Area under the curve
  - Test: 0.617106179584
  - Training: 0.775251922562




```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=18,
            max_features=None, max_leaf_nodes=None,
            min_impurity_split=1e-07, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            presort=False, random_state=1, splitter='best')
```

Applications & next steps

- Applications
  - Consumer profiling
  - Predict other information EG Education
- Next steps
  - Find variables that can render better salary predictions
  - ...