Qualogy - Data Science test case

Presentation

Jaap Broeders
06/05/2022

Table of contents

Case overview

‘Predicting preferred accommodation for a personalized travel recommendation system ‘

Use traveler and trip data to predict the preferred accommodation type for each traveler

Data importation methods

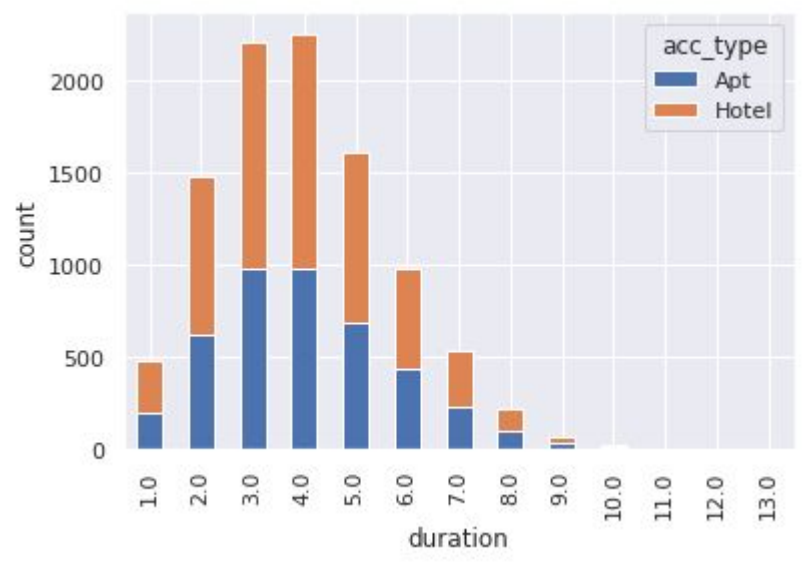
- Convert .txt data to an object which can be used for easy manipulation

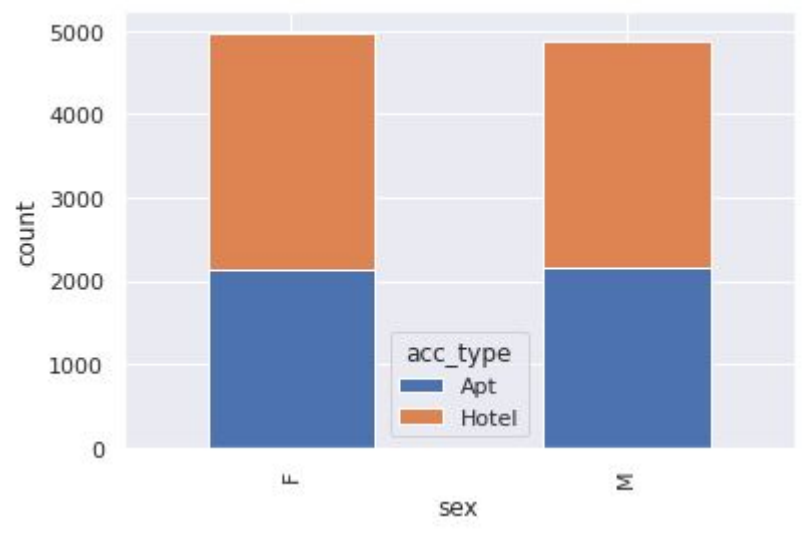
- Through general python string manipulation methods

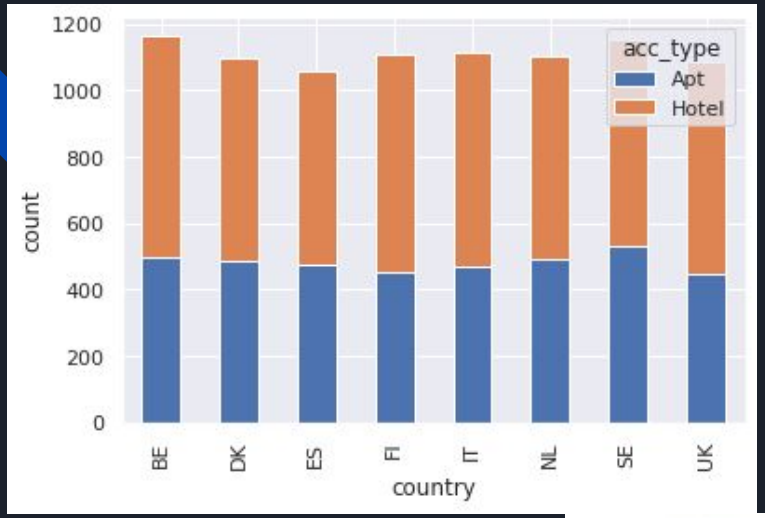
- Python libraries:
  - Pandas
  - Numpy
  - Re (regular expressions)

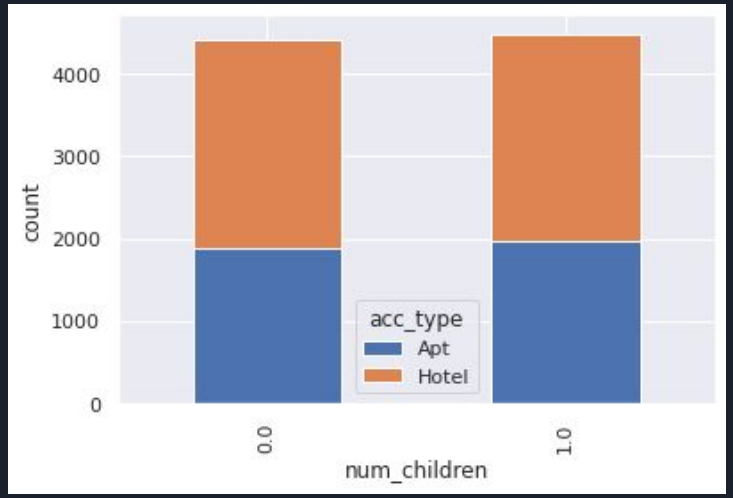
Exploratory Data Analysis

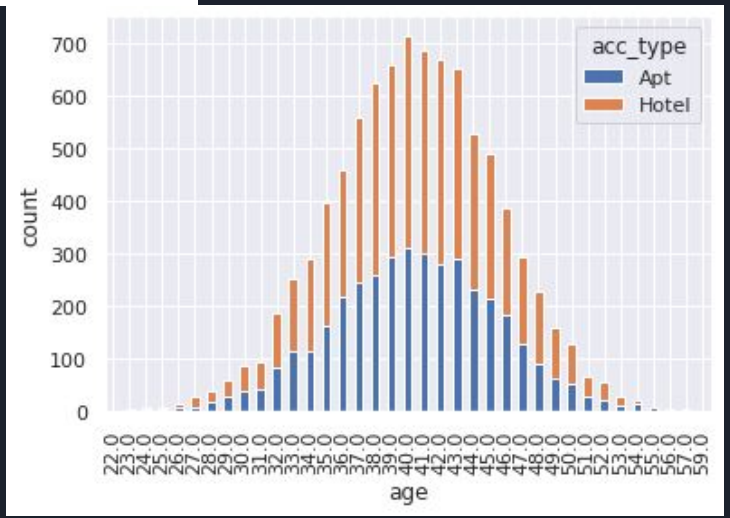
- Analyze and identify trends in the dataset
- Check data integrity
- Visualizations (binary target variable)
  - Stacked bar plots of feature frequencies related to the target variable
  - Confusion matrix
- Python libraries:
  - Pandas
  - Numpy
  - Matplotlib.pyplot
  - Seaborn


acc_type
Apt
Hotel
count
2000
1500
1000
500
0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0
duration


5000
4000
3000
2000
1000
0
count
acc_type
Apt
Hotel
F
M
sex


acc_type
Apt
Hotel
count
country
BE
DK
ES
FI
IT
NL
SE
UK
0
200
400
600
800
1000
1200


acc_type
Apt
Hotel
count
num_children
0.0
1.0
0
1000
2000
3000
4000


acc_type
Apt
Hotel
count
age
0
100
200
300
400
500
600
700

Correlation Matrix

	id	duration	sex	age	num_children	acc_type
id	1.000000	0.000769	0.013521	-0.021860	-0.007013	0.014468
duration	0.000769	1.000000	0.008632	0.017139	-0.000248	-0.003178
sex	0.013521	0.008632	1.000000	0.009047	0.020868	-0.014056
age	-0.021860	0.017139	0.009047	1.000000	-0.011178	-0.000013
num_children	-0.007013	-0.000248	0.020868	-0.011178	1.000000	-0.011502
acc_type	0.014468	-0.003178	-0.014056	-0.000013	-0.011502	1.000000

Data cleaning

- Identify and fix issues present within the dataset

- Verify correct data type for each feature in dataset
- Convert categorical binary features to numeric [0, 1]
- Add dummy variables for country feature
- Remove unnecessary features from dataset

- Python libraries:
  - Pandas

Data pre-processing

- Convert raw data into a representation suitable for application in ML models

- Impute missing values
  - Simple strategy: mean
  - Advanced possibility: Multivariate imputation
- Scale features
  - Strategy: Min-max scaling
  - Other possibilities: normalization/regularization algorithms
- Model input preparation
  - Separate training features X from target variable y
  - Split dataset into a training and test set

- Python libraries:
  - Pandas
  - sklearn

Modeling methods

- Random Forest (baseline)

- Artificial Neural Network

- Python libraries:
  - Numpy
  - Sklearn
  - Tensorflow/keras

Model evaluation methods

- Identify whether a model is adequate for the classification task

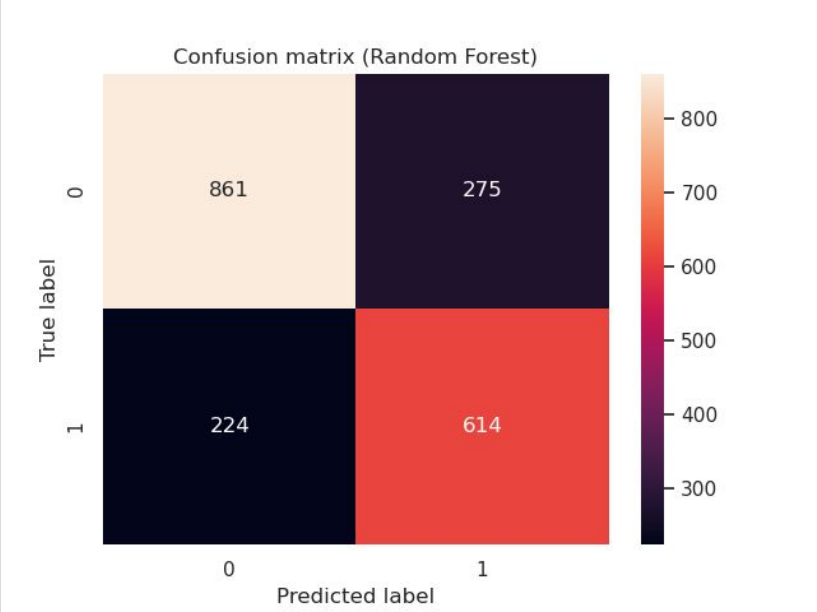
- K-fold cross validation

- Metrics (binary classification)
  - Accuracy, precision, recall, f1

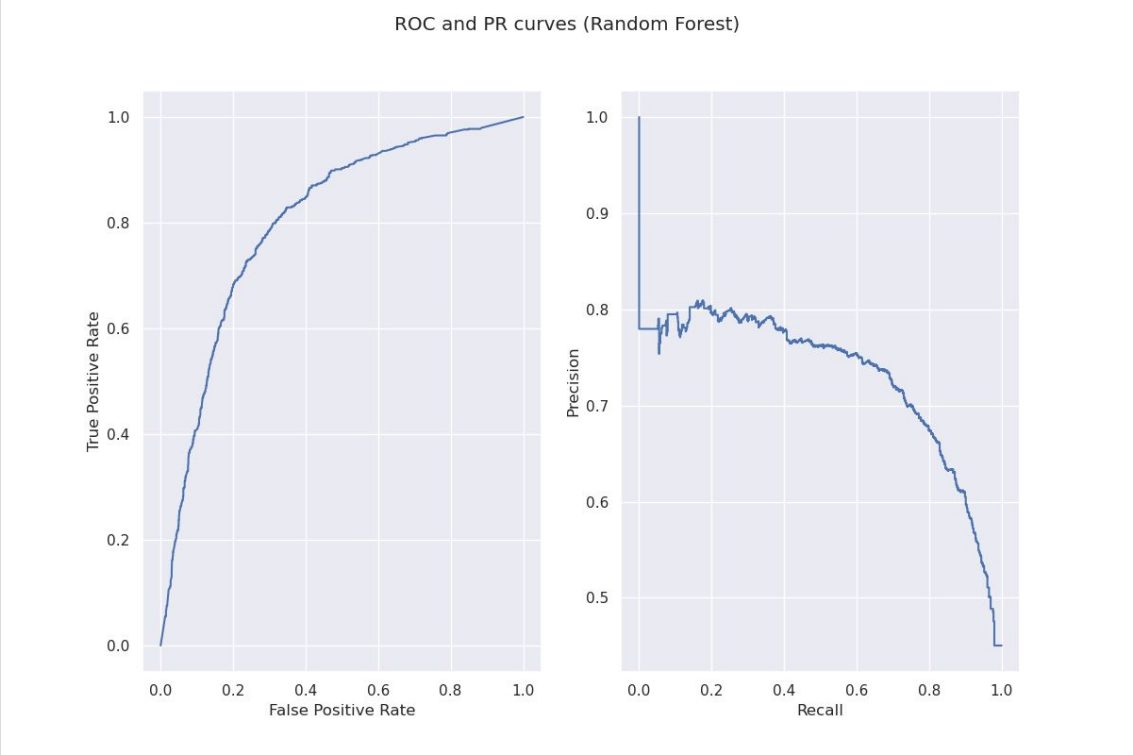
- Visualizations
  - ROC curve
  - Precision/Recall curve
  - Confusion matrix

Results - Random forest

```
MODEL: Random Forest
              precision    recall  f1-score   support

           0       0.79      0.76      0.78      1136
           1       0.69      0.73      0.71       838

    accuracy                           0.75      1974
   macro avg       0.74      0.75      0.74      1974
weighted avg       0.75      0.75      0.75      1974

QApplication: invalid style override passed, ignoring it.
Random Forest accuracy: 0.7404758566894206
```



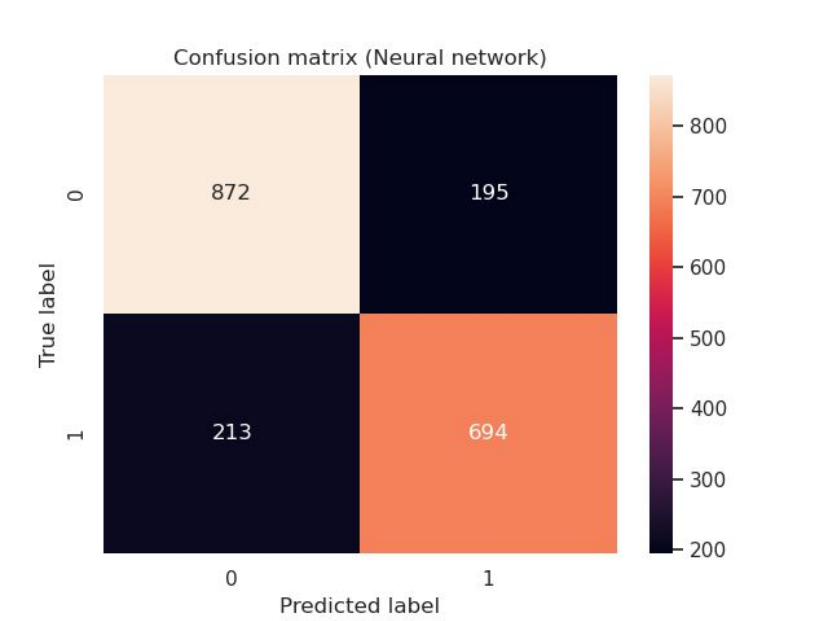
Results - Random forest



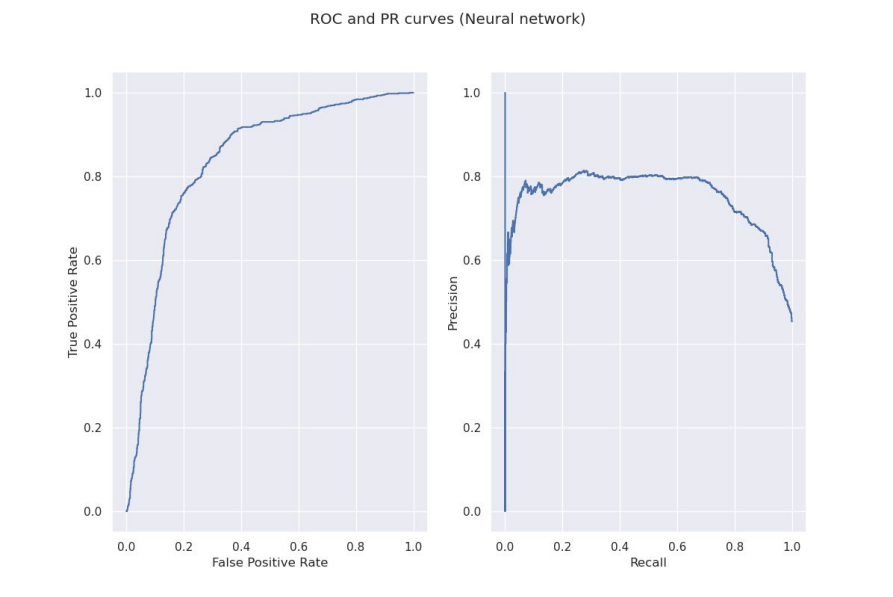
Results - Neural net

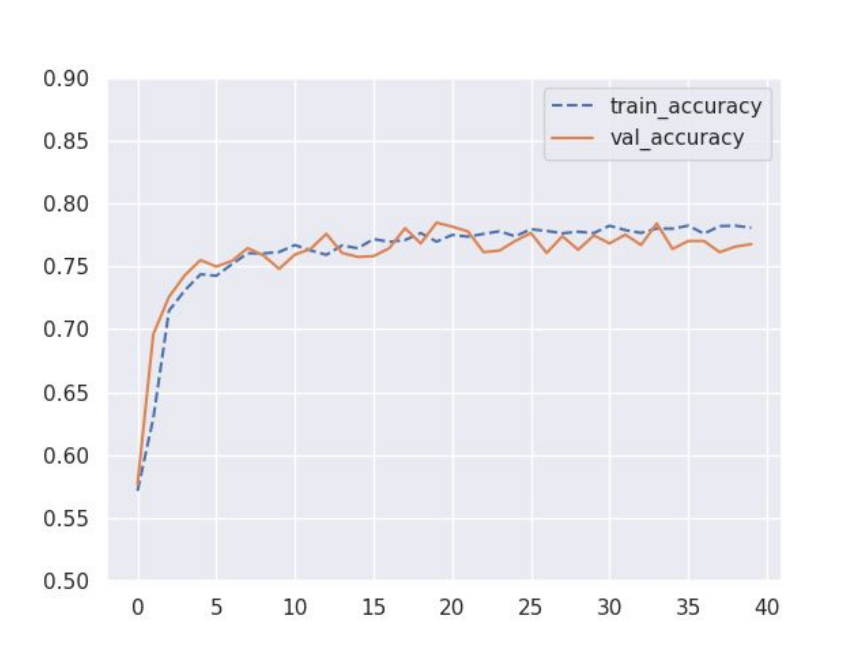
MODEL: Neural network

	precision	recall	f1-score	support
False	0.80	0.82	0.81	1067
True	0.78	0.77	0.77	907
accuracy			0.79	1974
macro avg	0.79	0.79	0.79	1974
weighted avg	0.79	0.79	0.79	1974



Results - Neural net





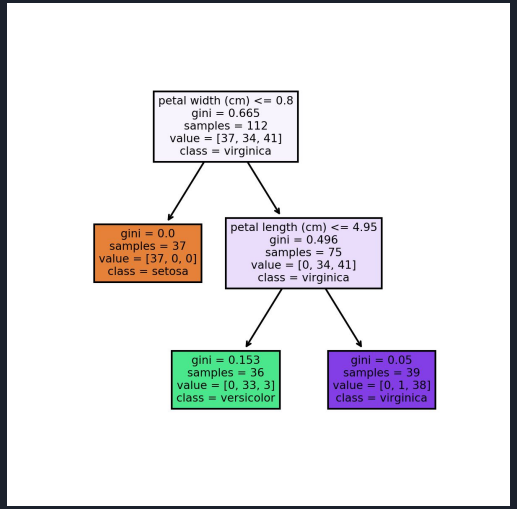
Future work

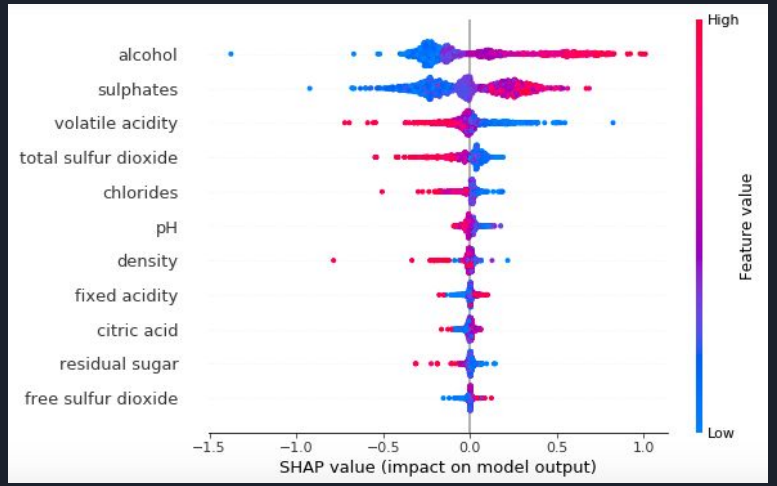
- Understand current models
  - Rf: visualize decision tree
  - ANN: SHAP

- Try other pre-processing methods
  - Feature engineering
  - Advanced imputation methods
  - Other scaling approaches

- Optimize ANN
  - Hyperparameters
  - Batch size, epochs

- Model deployment
  - AWS: Sagemaker
  - GCP: AI Platform





That was it!

Questions?