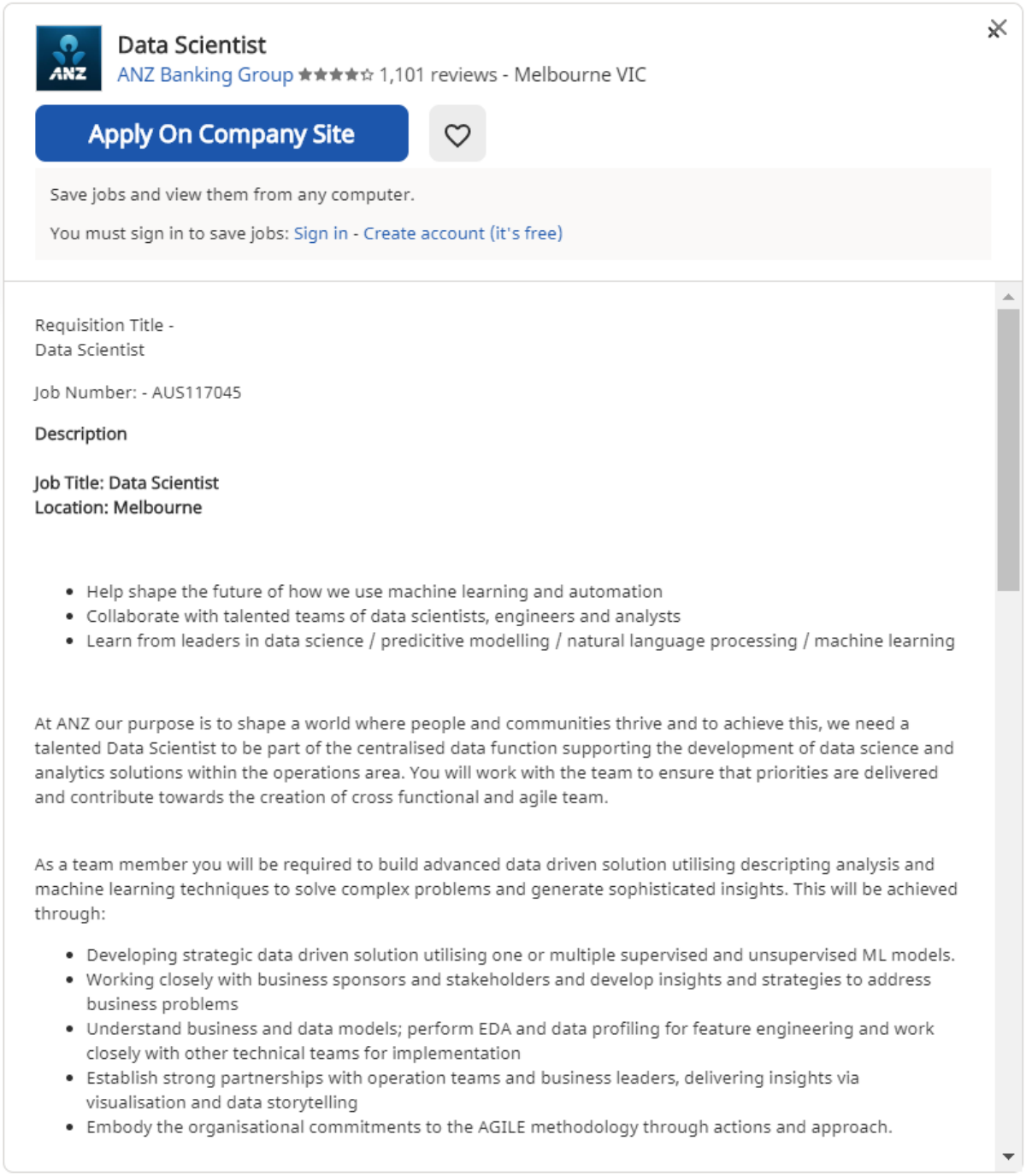
Part 1: Job Role

Find a job advertisement for a data science position that covers the type of data science role you see yourself doing.

a) Include a copy of the job description.



Data Scientist

ANZ Banking Group ★★★★☆ 1,101 reviews - Melbourne VIC

Apply On Company Site

Save jobs and view them from any computer.

You must sign in to save jobs: Sign in - Create account (it's free)

Requisition Title -
Data Scientist

Job Number: - AUS117045

Description

Job Title: Data Scientist
Location: Melbourne

- Help shape the future of how we use machine learning and automation
- Collaborate with talented teams of data scientists, engineers and analysts
- Learn from leaders in data science / predicitive modelling / natural language processing / machine learning

At ANZ our purpose is to shape a world where people and communities thrive and to achieve this, we need a talented Data Scientist to be part of the centralised data function supporting the development of data science and analytics solutions within the operations area. You will work with the team to ensure that priorities are delivered and contribute towards the creation of cross functional and agile team.

As a team member you will be required to build advanced data driven solution utilising descripting analysis and machine learning techniques to solve complex problems and generate sophisticated insights. This will be achieved through:

- Developing strategic data driven solution utilising one or multiple supervised and unsupervised ML models.
- Working closely with business sponsors and stakeholders and develop insights and strategies to address business problems
- Understand business and data models; perform EDA and data profiling for feature engineering and work closely with other technical teams for implementation
- Establish strong partnerships with operation teams and business leaders, delivering insights via visualisation and data storytelling
- Embody the organisational commitments to the AGILE methodology through actions and approach.

As a member of the Centralised Data Team, you should be forward thinking, work collaboratively and inspire others to ensure a strong and mature compliance culture.

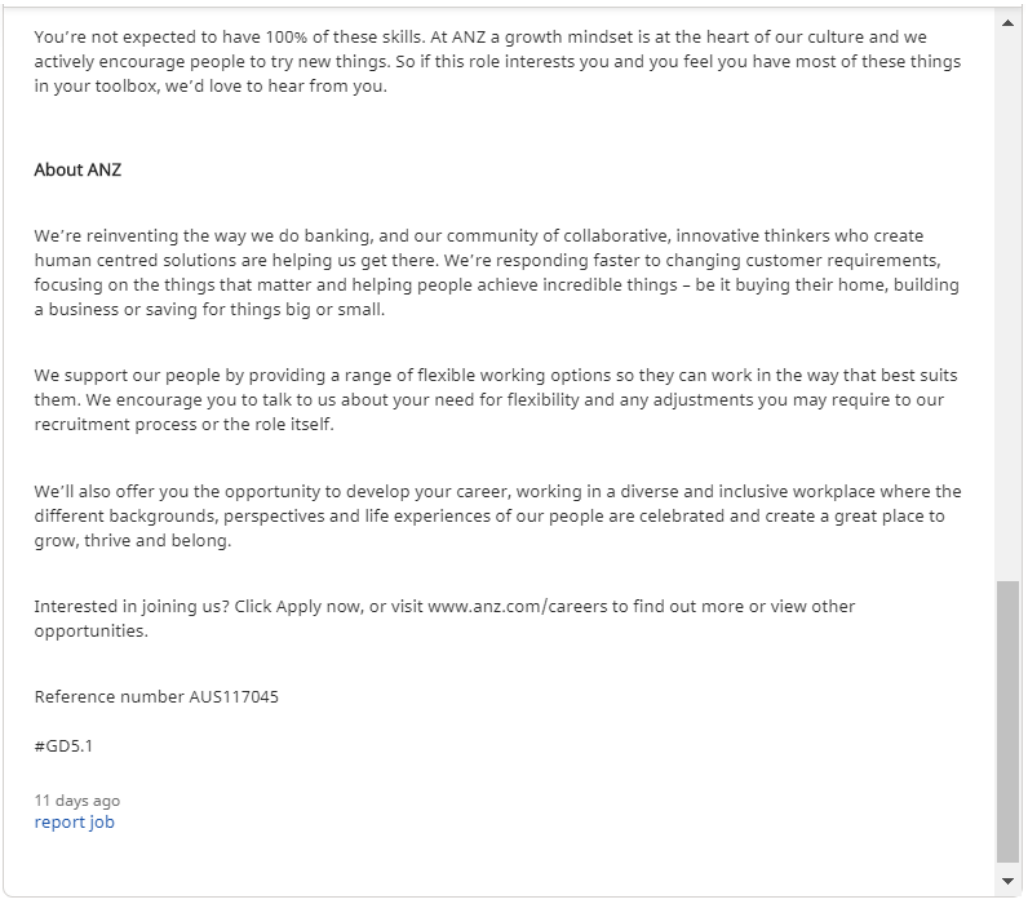
What a typical day might look like

- Attend stand-up to align progress on your individual tasks
- Support in delivering key analytics initiatives
- Work in collaborative teams to explore business problems
- Explore data on hand and identify data science opportunities
- Provide expert actionable advice on next steps to improve business bottlenecks
- Contributes to analytics communities, meetups and conferences to promote technology development culture and practices
- Speak to business stakeholders to sense check analysis results and workshop solutions

What will likely be in your toolkit?

- Tertiary qualified and experience working as a data scientist or analytics/modelling role
- Industry experience in machine learning model development and pipeline, tuning and deployment.
- Good working knowledge of data preparation and manipulation
- Solid understanding in underlying statistic and mathematic theory behind model used. (strong understanding in optimisation is favourable.)
- Strong programming experience with Python, R and SQL with the ability to articulate and demonstrate good coding practices
- Able to deliver consistent and accurate and repeatable analysis. Presenting current trends and emerging issues to key stakeholders
- Demonstrated effective deployment of models within a commercial environment with examples of improved performance, productivity, decision making and/or automation
- Ability to work independently and as a part of a team to drive initiatives in an agile delivery environment
- Continue to improve data science and business knowledge in operational risk and financial crime area.

You're not expected to have 100% of these skills. At ANZ a growth mindset is at the heart of our culture and we



You're not expected to have 100% of these skills. At ANZ a growth mindset is at the heart of our culture and we actively encourage people to try new things. So if this role interests you and you feel you have most of these things in your toolbox, we'd love to hear from you.

About ANZ

We're reinventing the way we do banking, and our community of collaborative, innovative thinkers who create human centred solutions are helping us get there. We're responding faster to changing customer requirements, focusing on the things that matter and helping people achieve incredible things – be it buying their home, building a business or saving for things big or small.

We support our people by providing a range of flexible working options so they can work in the way that best suits them. We encourage you to talk to us about your need for flexibility and any adjustments you may require to our recruitment process or the role itself.

We'll also offer you the opportunity to develop your career, working in a diverse and inclusive workplace where the different backgrounds, perspectives and life experiences of our people are celebrated and create a great place to grow, thrive and belong.

Interested in joining us? Click Apply now, or visit www.anz.com/careers to find out more or view other opportunities.

Reference number AUS117045

#GD5.1

11 days ago
report job

b) Which domain/field is the job position related to?

The job position is in ANZ Banking Group for Data Scientist which falls under Banking Domain.

c) Is the role looking for insights into the data or making predictions? Justify your answer.

The role which is for Data Scientist at ANZ Banking Group is looking for the Insights into the data by using various Data Mining techniques. As a team member it will be required to build advanced data driven solution utilising descripting analysis and machine learning techniques to solve complex problems and generate sophisticated insights. The role also includes having knowledge of Data mining techniques such as Classification and clustering. This can be achieved through developing strategic data driven solution utilising one or multiple supervised and unsupervised ML models. Working closely with business sponsors and stakeholders and develop insights and strategies to address business problems. Understand business and data models; perform EDA and data profiling for feature engineering and work closely with other technical teams for implementation. Establish strong partnerships with operation teams and business leaders, delivering insights via visualisation and data storytelling. The role requires using various machine learning models and presenting insights to stakeholders, from these it is clear that the role is for the insights into the data.

Part 2: Data Set

Find a public data set that could be construed as relevant in some way to the particular job ad you identified (for example, a data set that can be used for training and testing a binary classifier).

a) Include a URL and a brief description of the data set.

URL: - https://www.kaggle.com/janiobachmann/bank-marketing-dataset

Description: -This is the classic bank marketing dataset uploaded originally in the UCI Machine Learning Repository. The dataset gives you information about a marketing campaign of a financial institution in which we have to analyze in order to find out customer segments, using data for customers, who subscribed to term deposit. This helps to identify the profile of a customer, who is more likely to acquire the product and develop more targeted marketing campaigns.

This dataset contains banking marketing campaign data and we can use it to optimize marketing campaigns to attract more customers to term deposit subscription.

A Term deposit is a deposit that a bank or a financial institution offers with a fixed rate (often better than just opening deposit account) in which your money will be returned back at a specific maturity time.

Attributes Information:

About client features:
-age
-job (type of job)
-marital (marital status)
-education
-default (has credit in default? 'no', 'yes' , 'unknown')
-balance
-housing (has housing loan? 'no', 'yes', 'unknown')
-loan (has personal loan? 'no', 'yes', 'unknown')

About previous contacts/campaign:
-pdays (number of days that passed by after the client was last contacted from a previous campaign, 999 means client was not previously contacted)
-previous (number of contacts performed before this campaign and for this client)
-poutcome (outcome of the previous marketing campaign: 'failure', 'nonexistent', 'success')

About current campaign:
-contact (communication type: 'cellular', 'telephone')
-duration (last contact duration, in seconds) Important note: the duration attribute highly affects the output target (e.g., if duration=0 then y='no').
-day (last contact day of the week)
-month (last contact month of year)
-campaign (number of contacts performed during this campaign and for this client)
-deposit (has the client subscribed a term deposit? 'yes', 'no')

b) Very briefly describe why you chose it (e.g. how it relates to the data science position).

The ANZ Banking Group is a Banking company which helps to create human centered solutions and are responding faster to changing customer requirements, focusing on the things that matter and helping people achieve incredible things be it buying their home, building a business or saving for things big or small. The ANZ Banking Group would have a database which would include all the information related to customers of bank such as customers age, job education, marital status, housing loan, balance and deposit and more bank related information. The Company may also have database such as customer's bank accounts, credit cards, Personal Loans, Insurance and other bank details. The Dataset which I have chosen is Bank Marketing database and which can be closely related to ANZ Banking Group. In this Dataset, as the data science position I can build advanced data driven solution utilising descripting analysis and machine learning techniques to solve complex problems and generate sophisticated insights. Here the Bank Marketing Dataset is used to optimize marketing campaigns to attract more customers to term deposit subscription. However, the dataset is closely related to Banking and the ANZ Banking Group can also have the same dataset related to bank Marketing. Therefore the reason to work on this dataset will be better as it is much related with the job position as Data Science for ANZ Banking Group.

Part 3: Experiment

Use two machine learning algorithms (e.g., including decision trees, kNN, SVM, or neural networks) to gain insight into the data. One of the machine learning models has to be different from the ones used in the Practical Data Science course.

a) What insights do the machine learning algorithms give you?

The algorithms used were KNN Classifier and Random Forest Classifier to gain the insights from the Bank Marketing dataset. KNN Classifier and Random forest Classifier are the examples of supervised machine learning. It depends on input data to learn a function that produces appropriate output for unlabelled data. This dataset contains banking marketing campaign data and we can use it to optimize marketing campaigns to attract more customers to term deposit subscription. The dataset had attribute as "deposit" which has binary values as yes or no and from this we get to know that whether to give that customer the deposit or not. The dataset is split into Training data which is 70% and testing data which is 30%. Splitting is done to get better indication of the models performance on unseen data. The accuracy increases when the value of k is small. The smaller k=1 can result in 100% accuracy and can also cause overfitting. Therefore, K=5 is the best value for KNN.

KNN:-K Nearest Neighbour algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. In KNN, the model structure is determined from the data

without making any assumptions on underlying distributions. K nearest neighbours is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions)

From KNN Classifier we get to know whether the person is able to get deposit or not and also we can know that whether a person is fit for loan approval or not. KNN algorithm can help when calculating an individual's credit score by comparing it with persons with similar traits.

Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision trees. It can be used for both classification and regression tasks. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. The random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. Random Forest algorithm can be used to understand the behaviour of the customers for eg. A customer can be given loan based on the customer account balance, previous customer transactions and also based on the customer's previous savings (fixed deposit)

KNN Classifier had an accuracy of 77% while random forest had an accuracy of 84%. We see that Random forest has performed well in comparison to KNN.

b) Compare the predictive power of the models produced by the two algorithms and visualise their effectiveness. Clearly justify the evaluation metrics you used to compare the effectiveness of your the models.

Confusion matrix: - A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm. It allows easy identification of confusion between classes e.g. one class is commonly mislabelled as the other. Most performance measures are computed from the confusion matrix.

The main purpose of a confusion matrix is to see how our model is performing when it comes to classifying potential clients that are likely to subscribe to a term deposit. We will see in the confusion matrix four terms the True Positives, False Positives, True Negatives and False Negatives.

- Classification error rate: the percentage of observations in the test data that the model mislabelled.

TN-True Negative (Top-Left Square): This is the number of correctly classifications of the "No" class or potential clients that are not willing to subscribe a term deposit.

TP-True Positive (Bottom-Right Square): This is the number of correctly classifications of the "Yes" class or potential clients that are willing to subscribe a term deposit.

FN-False Negative (Top-Right Square): This is the number of incorrectly classifications of the "No" class or potential clients that are not willing to subscribe a term deposit.

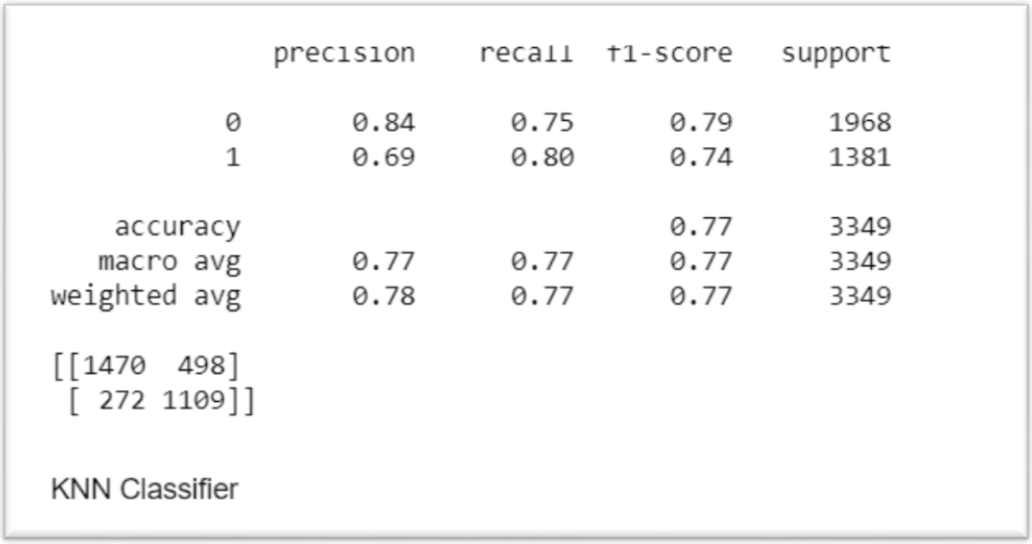
FP-False Positive (Bottom-Left Square**):** This is the number of incorrectly classifications of the "Yes" class or potential clients that are willing to subscribe a term deposit.

- Precision: The precision is the ratio tp/ (tp+fp) where tp is the number of true positives and fp is the number of false positives i.e the ability of the classifier not to label as positive a sample that is negative. Precision means how correct the prediction of our model that has actual label as “Yes”

- Recall: The recall is the ratio tp / (tp + fn) where tp is the number of true positives and fn is the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples. So how many "Yes" labels does our model detect is done by recall.

- F1-score: The F1 score can be interpreted as weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is: F1 = 2 (precision recall) / (precision + recall)

- Support: Support is the number of samples of the true response that lie in the class.

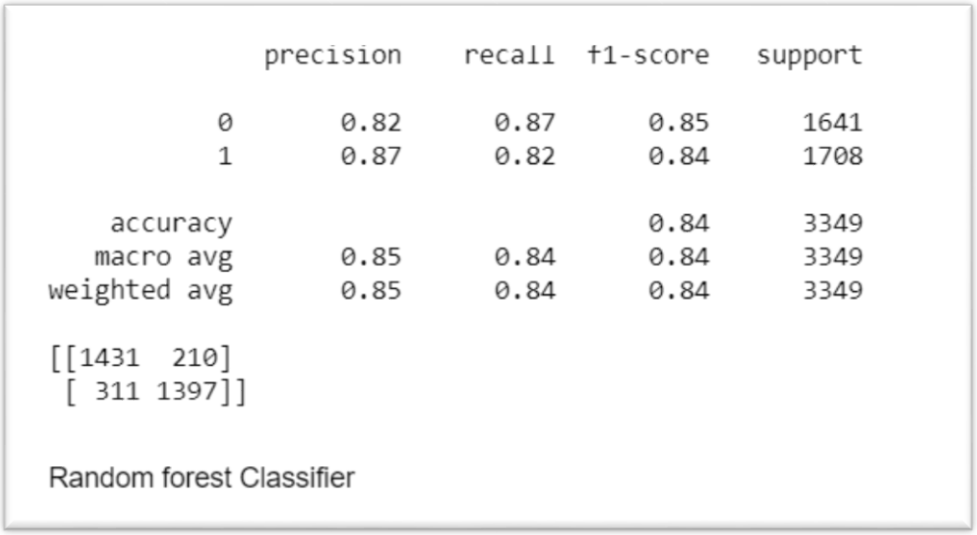


	precision	recall	f1-score	support
0	0.84	0.75	0.79	1968
1	0.69	0.80	0.74	1381
accuracy			0.77	3349
macro avg	0.77	0.77	0.77	3349
weighted avg	0.78	0.77	0.77	3349

```
[[1470  498]
 [ 272 1109]]
```

KNN Classifier

The above image shows the Classification report for the KNN Classifier. In this algorithm both the labels were identified. A Classification report is used to measure the quality of predictions from a classification algorithm. In the above classification report we can say that there are 1968 instances of 0 label and 1381 instances of 1 label which is for support. Support is the number of samples of the true response that lie in the class. Precision means how correct the prediction of our model that has actual label as “Yes”. In this case there are 0.84 for 0 label and 0.69 for 1 label. So how many "Yes" labels does our model detect is done by recall. In this case 0.75 is for 0 label and 0.80 is for 1 label.F1 score is the weighted average of the precision and recall. The better the F1 score the better is the performance of the model. For the above confusion matrix for KNN Classifier we see that diagonal elements are those which are correctly identified labels. In these case 1470 and 1109 are correctly identified labels and 498 and 272 are incorrectly identified.



	precision	recall	f1-score	support
0	0.82	0.87	0.85	1641
1	0.87	0.82	0.84	1708
accuracy			0.84	3349
macro avg	0.85	0.84	0.84	3349
weighted avg	0.85	0.84	0.84	3349

```
[[1431  210]
 [ 311 1397]]
```

Random forest Classifier

The image shows the classification report for Random forest Classifier. In the above classification report we can say that there are 1641 instances of 0 label and 1708 instances of 1 label which is for support. Support is the number of samples of the true response that lie in the class. Precision means how correct the prediction of our model that has actual label as “Yes”. In this case there are 0.82 for 0 label and 0.87 for 1 label. So how many "Yes" labels does our model detect is done by recall. In this case 0.87 is for 0 label and 0.82 is for 1 label.F1 score is the weighted average of the precision and recall. The better the F1 score the better is the performance of the model. For the above confusion matrix for Random Forest algorithm we see that diagonal elements are those which are correctly identified labels. In these case 1431 and 1397 are correctly identified labels and 210 and 311 are incorrectly identified.

By comparing both the models we see that F1 score of Random Forest classifier is better which means that the model performs better than KNN classifier. Therefore, Random Forest has better F1 score, accuracy and confusion matrix in compare to KNN Classifier. By comparing both the models we can say that Random Forest Algorithm performs better than KNN Algorithm.

References

[1]"Bank Marketing Campaign || Opening a Term Deposit", *Kaggle.com*, 2020. [Online]. Available: https://www.kaggle.com/janiobachmann/bank-marketing-campaign-opening-a-term-deposit. [Accessed: 03- Aug- 2020].

[2]"Bank Marketing", *Kaggle.com*, 2020. [Online]. Available: https://www.kaggle.com/rishabdhar1619/bank-marketing. [Accessed: 03- Aug- 2020].

[3]"Random Forest—A powerful Ensemble Learning algorithm", *Medium*, 2020. [Online]. Available: https://towardsdatascience.com/random-forest-a-powerful-ensemble-learning-algorithm-2bf132ba639d. [Accessed: 03- Aug- 2020].