Machine Learning Test

1. Airline Use Case

Suppose we build a recommendation engine for a budget airline that predicts online purchases of pre-booked items for each passenger, such as priority boarding, extra leg room, exact seating, and food and beverages.

**Solutions :-**

Question 1.1

There is a confusion at the client about how to set up the machine learning task. What are the training examples (X) and what are the predicted outcomes (y)? What kind of algorithm to use? What should be the evaluation metric of the model? Please share your suggestions with them for each question.

Solution 1.1:

For the above problem our Training example (X) should contains passenger’s implicit information like history of purchases as well as passenger’s explicit information like rating on food and beverages given by passenger.

The outcome (Y) should be the Item id that passenger will likely to buy.

Matrix Factorization would be my first priority in building a recommendation engine and on top of this we can use Recall or Mean Reciprocal Rank as our evaluation metrics.

### Question 1.2

We know that the popularity of the products are varying. E.g. purchase rate for priority boarding is 20 % while for food is 2 %. How would this influence the recommendations? Do we need to handle it somehow?

Solution 1.2:

If the purchase rate of boarding is more than our dataset has more information about boarding than food, therefore our recommendation system will be more likely to predict boarding more often.

What we can do is try some sampling method in data to remove this imbalanced classification, and then build recommendation on top of that.

The other method could be that you build different recommendation engine for different category which might be computationally expensive but could yield better results.

### Question 1.3

We settled to use one year’s data of online pre-booked purchase behavior for model training, which we split into 70% training and 30% evaluation sets randomly. Our final model is ready and it performs well on both sets. The plan is to retrain the model (no hyperparameter-tuning, just re-run) every day at 1 am based on data of the previous 30 days.

A data scientist from the client’s team expresses concerns that the production system will not perform as well as indicated by our training setup. Is this concern valid? How would you address his concern? Write an email to him.

Solution 1.3:

Sir,

This is surely a valid concern, it is a very bad idea to just re run the same model without seeing the distributions of the new data. Our new data might be biased which could affect our previous approach and performance using older data.

New data should have similar past distributions to give same results.

Regards,

Rohit Chandravanshi

## **2. ML methodology**

### Question 2.1

A new classifier model identifies bad bonds in the financial market for a hedge fund. Bad bonds can have devastating effects and must be avoided in the portfolio. 0.01% of all bonds fall into this category and our model has an accuracy of 99.99%. Is this ML model doing a good job? Why?

Solution 2.1:

Here Accuracy is not the right metric to decide the performance of our model. Since this is a problem of imbalanced classification, so our model will likely to give prediction on the class that has very high weights in the data.

So we have to use some other metrics like F1 score to correctly measure the performance of our model.

Fill in the empty confusion matrix below with a possible concrete outcome if there are 100,000 bonds in the market.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual | |
|  |  | Bad | Not bad |
| Predicted | Bad | 1000 | 100 |
| Not bad | 900 | 98000 |

### Question 2.2

On a logistic regression model with binary outcome in {0,1} that is optimized with stochastic gradient descent you have to tune hyperparameters

* learning rate
* L2 regularization
* batch size
* threshold value: the predicted probability above which we assign 1

Choose 3 metrics that you can use to compare the trained model and decide which one is the best for this use case. Explain why.

Solution 2.2

Three metrics that I have chosen for the above case are as:

1. PRECISION:- Here in the problem number of negative samples are very large hence precision would be a great measure for deciding the model performance.
2. F1 Score – F1 score is a good measure for imbalanced classification because it takes both false negative and false positive into account.
3. ROC-AUC – By using this curve we can estimate a better threshold value for minority classes.

From above metrics I would have chosen ROC-AUC as our confusion metric have some False positives too and ROC only accounts true positive rates and false positive rate by which we can yield a better threshold for minority classes.

## 3. Math

### Question 3.1

The matrix below shows the probability that you are in a state today given we know your state from yesterday. There are two states: each day either you read or you train.

When you read one day you are very likely to continue the book the following day.

When you train you decide with a coin flip whether to go out and train again the next day.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Today | |
|  |  | Reading | Training |
| Yesterday | Reading | 0.9 | 0.1 |
| Training | 0.5 | 0.5 |

Which mathematics concept would you use to calculate the probability of training at any given day? (The probability of training after an infinite number of days?) You are not required to calculate it.

Solution 3.1:

To calculate probability of training at a given day, we can use conditional probability of two events of coin flipping and training together.

The probability of training after an infinite number of days would be equal to 1.