Instacart Market Basket Analysis

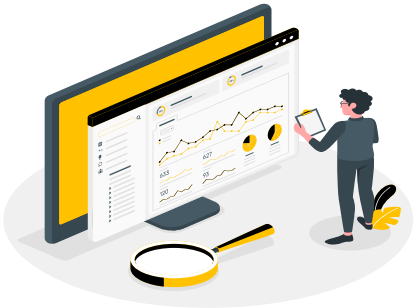


Table of Contents

01 Project Background & Objectives

Background

- Data comes from Kaggle

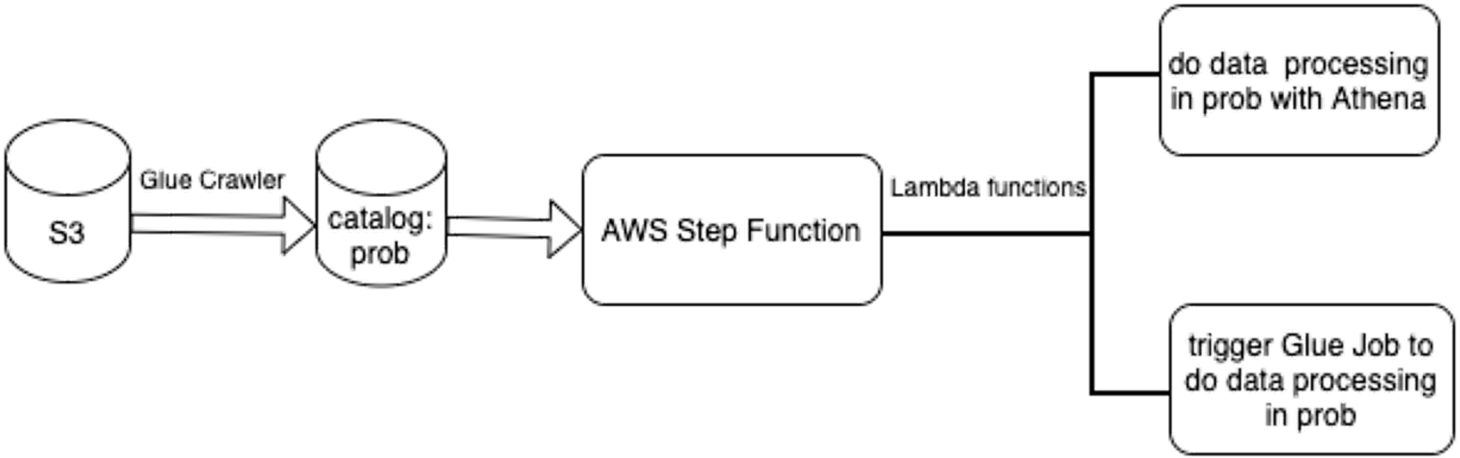
- To provide a delightful shopping experience by using customer orders over time to predict which previously purchased products will be in a user's next order

Instacart Market

Objectives

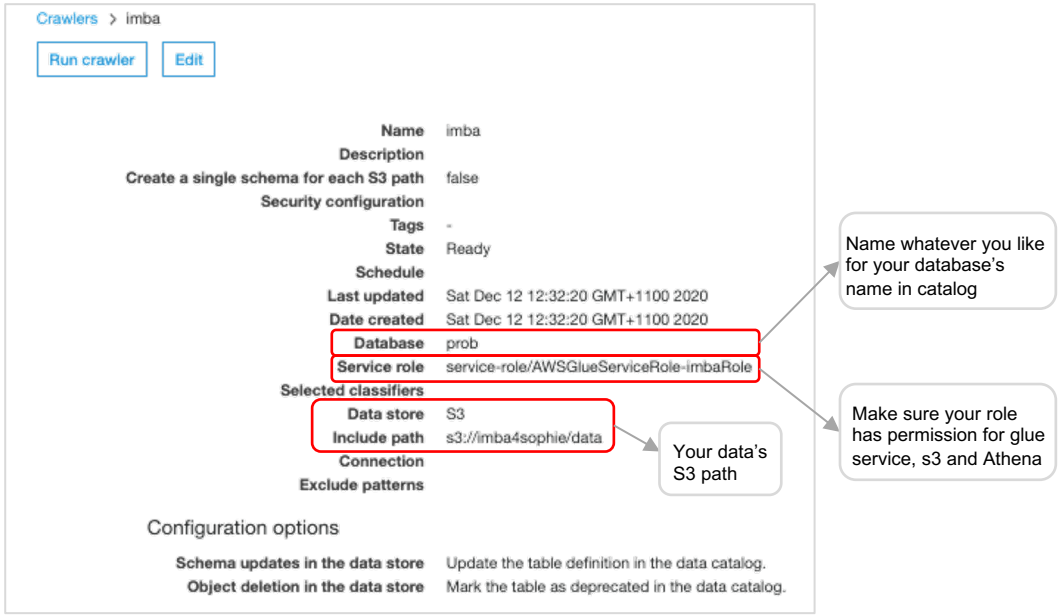
- Build automated ETL pipeline to process big data
- Build model to do the prediction
- Build the user interface

02 Automated ETL pipeline to process big data



02 Automated ETL pipeline to process big data

Glue Crawler



02 Automated ETL pipeline to process big data

Step Function

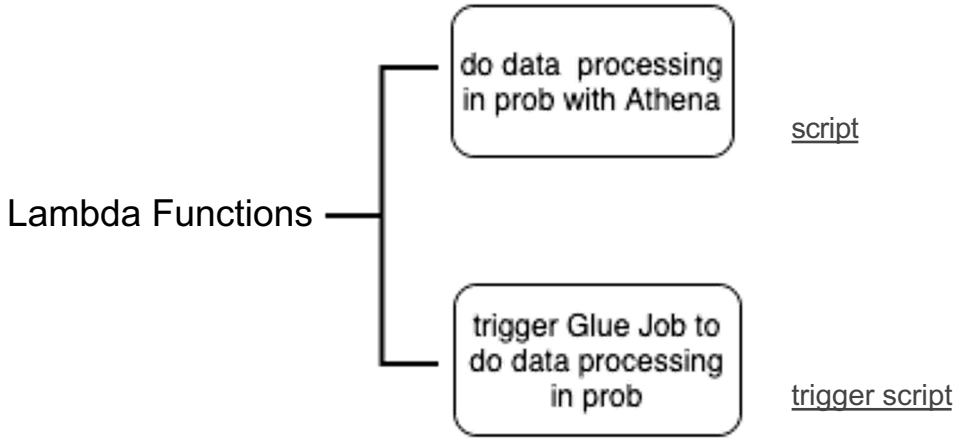
- Create a state machine with script in definition and execution input as below figure showing
- Make sure it has Lambda permission

Details | **Execution input** | Execution output | Definition

```
{
  "bucket": "imba4sophie",
  "prefix": "features/",
  "database": "prob",
  "query_output": "s3://imba4sophie/query_results/"
}
```

Replace with your catalog and S3 bucket for data and output

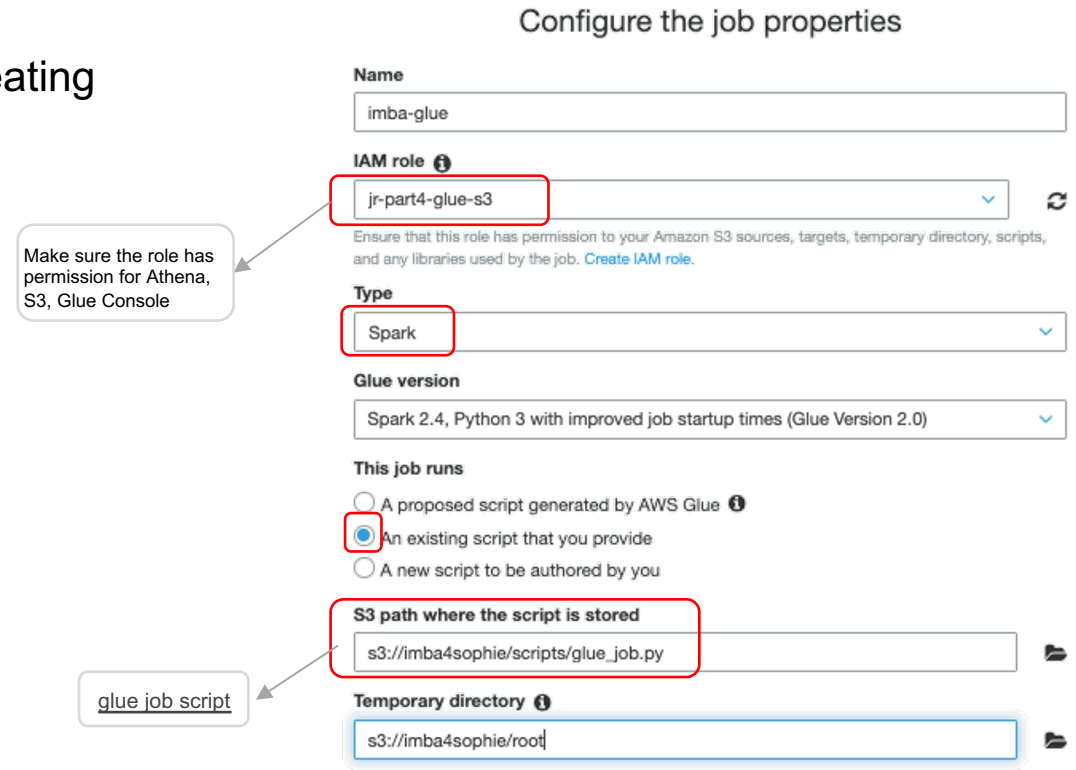
02 Automated ETL pipeline to process big data



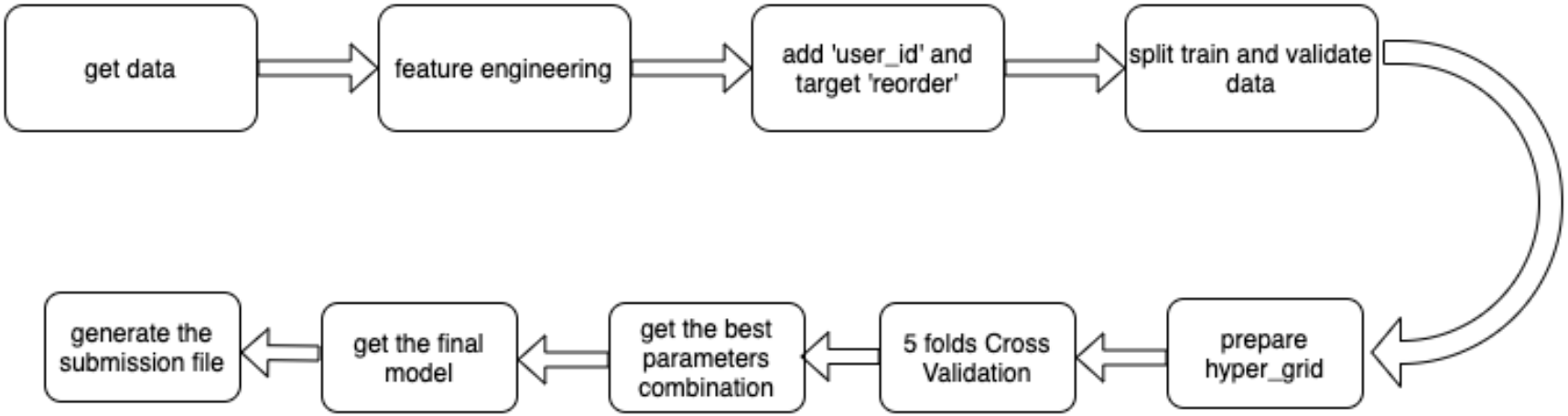
Note: make sure your lambda function role has permission for Glue Console, Athena and Lambda

02 Automated ETL pipeline to process big data

Glue Job Creating



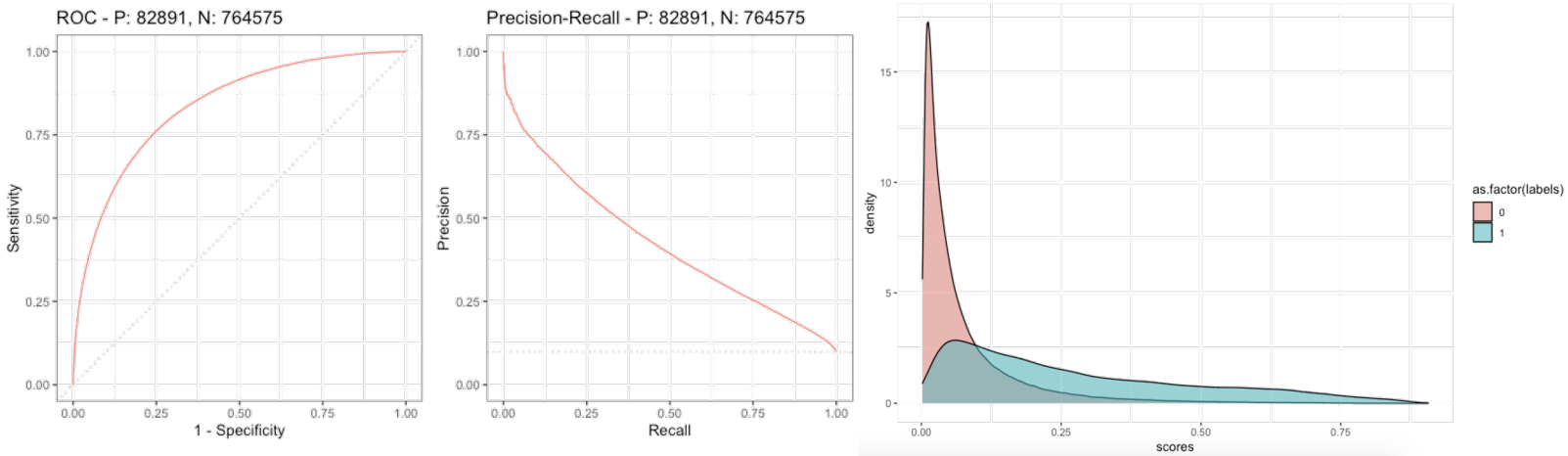
03 Data Modelling



03 Data Modelling

Products Reordered Prediction R script

- The goal is to predict if the purchased product will be ordered again
- The model was built using Xgboost in R
- The model achieved a test AUC of 0.832
- R Libraries used: ProjectTemplate, tidyverse, xgboost, pROC, precrec



04 Deployment

To be added…