Columbia Sportswear Data Science Predictive Modeling Assessment

Introduction

At Columbia, our data science team is continually striving to improve our predictive models. Many of the problems we aim to solve require developing probabilistic models that quantify the chance that an event will or will not happen. In this exercise, you will be tasked with fitting a predictive model which assigns a probability that a product from an order will be returned by the purchaser.

Data

Included are two files: train.csv and test.csv. Each contains synthetically generated transactions from an imaginary clothing retailer located in the USA. Each row represents a line item from a specific online order. A description of the columns is provided:

ID	The unique ID of the transaction
OrderID	The unique ID of the order
CustomerID	The unique ID of the customer
CustomerState	The US territory of the customer
CustomerBirthDate	The birth date of the customer
OrderDate	The date the order was placed
ProductDepartment	The retail department of the product
ProductSize	The size of the product
ProductSize	The manufacturing cost of the product
DiscountPct	The discount that was applied to the transaction
PurchasePrice	The sale amount paid by the customer
Returned	Whether the product was eventually returned

Regarding external data sources (e.g., weather, economic indicators, etc.), the data were synthetically generated so it is unlikely that anything external to the dataset would be related to the response variable. Everything you need to make a good model is already in here. Each product that is sold by this retailer is unique on ProductDepartment, ProductCost and MSRP. MSRP can be recovered using DiscountPct and PurchasePrice columns.

Project

Using the transactions in the train.csv file, you will train a model that will calculate the probability that a transaction results in a return. You will score the rows in the test.csv file using your model. Your code used to produce the submission may be in R or Python, but Python is preferred. Your submission file should be a csv in the following format:

ID	Prediction
e72r-4033-a4cf	0.345525
cv70-4e4a-bq0e	0.662342
734e-4ak1-ad2e	0.219004

There should be one prediction for every transaction in the test.csv file. You will email your submission to an address which will be provided.

Additionally, please provide the code you used to produce the submission, as well as a short write up explaining your approach and methodology.

Evaluation

Your submission will be evaluated on ROC AUC (area under the receiver operating characteristic curve) using the predicted probability and the observed target. If your submission meets the ROC AUC score threshold, an additional evaluation will be done using Brier score to test calibration.