ASOS Data Scientist Exercise

Please review the information below and then prepare answers to the questions posed in the next section. You will be required to present your answers at an interview, so please prepare them in an appropriate format, including any solution diagrams, visualizations, models and code. You must submit your presentation in advance of your interview for review. You will have approximately 1.5 hours to present your answers, including discussion time.

This exercise is designed to be **very representative** of the types of problems you would encounter at ASOS, so we hope you enjoy the challenge!

ASOS Data Source Overview

[This has been simplified for commercial sensitivity and for the purpose of the exercise] ASOS has the following sources of customer centric data relevant to this set of tasks:

A CRM System that contains customer level information. Attributes include: order detail (date/time of order, basket size, quantity and value of SKUs within the basket, information on returned products), SKU detail (brand, category, size, colour, price), campaign information (receipts, opens, click-through of email campaigns) and opt-in information (whether the customer has opted in to receive marketing emails and when they subscribed or unsubscribed). **The interface to the system** is via a web based GUI or direct SQL queries.

A Web Analytics System that contains aggregated web traffic information, such as total counts of visitor and visit numbers, geography, pages visited, traffic sources, entry and exit rates, conversion funnels **The interface to the system** is via a web based GUI or REST API returning HTML, JSON or XML.

A Cloud Based Log Data Collection Solution that currently contains raw clickstream data from the Web Analytics system held at the lowest possible granularity. Every event on the website is captured and includes detail on the event, the session and the visitor. This data is the order of a billion rows a month. It is possible to upload raw data from other systems mentioned into this system**. The interface to the system** is programmatic.

A Social Media monitoring system that enable us to capture all posts and their metadata from a variety of social media and blogging sites that match a filter set of keywords that we define. For example, we can capture every mention of the ASOS and

key competitor brands on Twitter. **The interface to the system** is through a web based GUI or API.

A Customer Care system that enable us to capture all semantic data relating to customer's request for assistance and any resulting outcome of this contact. For example, we can capture every time a customer contacts us about a late delivery, or any issues related to their order or issues they might be having with our website or mobile app. **The interface to the system** is through a web based GUI or API.

A transactional database that captures all purchases and returns information related to customer orders. **The interface to the system** is both programmatic and via a GUI.

Task 1

ASOS would like to understand more about our customer's behavior, and in particular the underlying factors that are predictive of customer churn. We would like you to create at least two customer churn models from the data provided using the following guidelines:

1. The 2+ models created should be from different modeling approaches (eg: logistic regression, decision trees, SVMs etc).
2. Compare the performance of each modeling approach in addition to the pros and cons of each approach given the problem and data.
3. Use a code repository to check in your code during the development process.
4. Use a programming language and software package of your choice.

In addition to these guidelines, outline and explain the following:

1. The methodology you implemented.
2. Any simplifying assumptions made.
3. Your feature selection / creation process.
4. How you would improve the models if you had access to more data sources or additional time.
5. The tools you would require to build test and run this model at scale.

Task 2

Given what you know about the ASOS business, list three other examples where you believe you could either create incremental improvement in or entirely new value (either financial or in terms of customer experience) by applying your skills to the varied and rich data sources we have.

For each example list the improvement you believe you could deliver and a brief overview of your approach. (1 slide)

Data provided

You have been provided with four (4) simplified and anonymised tab delimited datasets that cover several years of activity for roughly 500k customers. The datasets are:

- Customer Demographics
- Receipt transactions
- Returns transactions
- Websession summaries (simplified weblog data)

Customer Demographics: (23MB)
Schema:
customerId2 int,
churnlabel int, (1 active customer, 2 churned customer)
gender string,
shippingCountry string,
dateCreated string,
yearOfBirth int,
premier int (1 Pending, 2 Active - NotUsed, 3 Active - Used, 4 Cancelled, 5 Lapsed, 6 Dormant)

Receipt transactions: (391MB)
Schema:
customerId2 int,
productId int,
divisionId int, (4 Men's Outlet, 5 Menswear, 6 Women's Outlet, 7 Womenswear)
sourceId int, (1 Full price purchase, 2 Discount code purchase, 3 Sales purchase, 4 other purchase, 10 returns)
itemQty int,
signalDate string,
receiptId int,
price double

Return transactions: (70MB)
Schema:
customerId2 int,
productId int,

divisionId int, (4 Men's Outlet, 5 Menswear, 6 Women's Outlet, 7 Womenswear)
sourceId int, (1 Full price purchase, 2 Discount code purchase, 3 Sales purchase, 4 other purchase, 10 returns)
itemQty int,
signalDate string,
receiptId int,
returnId int,
returnAction string,
returnReason string

WebSession Summaries: (2.1GB)
Schema:
customerId2 int,
country string,
startTime string,
site string,
pageViewCount int,
nonPageViewEventsCount int,
userAgent string,
screenResolution string,
browserSize string,
productViewCount int,
productViewsDistinctCount int,
productsAddedToBagCount int,
productsSavedForLaterFromProductPageCount int,
productsSavedForLaterFromCategoryPageCount int,
productsPurchasedDistinctCount int,
productsPurchasedTotalCount int

Accessing the Data

The data is located in Azure blob storage.

Windows

Under Windows the easiest way to access the data is using the free cloudberry tool (http://www.cloudberrylab.com/free-microsoft-azure-explorer.aspx).

To connect to the storage location, you would need the following credentials:

Account name: asosdsrecruiting

Access key:
QWGmgaTqn+Q+0YLT/47zM79HfNbSK0lX98wvoKqdCuGi7Q4inQGMKCpe+seF5qQ1xYEK
j+6E5Zp6ZRELGYjSOg==

The data locations are:

Customer Demographics:

Location: wasb://recruitingdata@asosdsrecruiting.blob.core.windows.net/customer

Receipts:

Location: wasb://recruitingdata@asosdsrecruiting.blob.core.windows.net/receipts

Returns:

Location: wasb://recruitingdata@asosdsrecruiting.blob.core.windows.net/returns

WebSession Summaries

Location:
wasb://recruitingdata@asosdsrecruiting.blob.core.windows.net/sessionsummary

Mac

Under mac the easiest way to access the data is using the free Cyberduck tool (https://cyberduck.io).

To connect to the storage location, you would need the following credentials:

Server: asosdsrecruiting.blob.core.windows.net
Username: asosdsrecruiting
Password:
QWGmgaTqn+Q+0YLT/47zM79HfNbSK0lX98wvoKqdCuGi7Q4inQGMKCpe+seF5qQ1xYEK
j+6E5Zp6ZRELGYjSOg==

The data locations are:

Customer Demographics:

Location: recruitingdata/customer

Receipts:

Location: recruitingdata/receipts

Returns:

Location: recruitingdata /returns

WebSession Summaries

Location: recruitingdata /sessionsummary