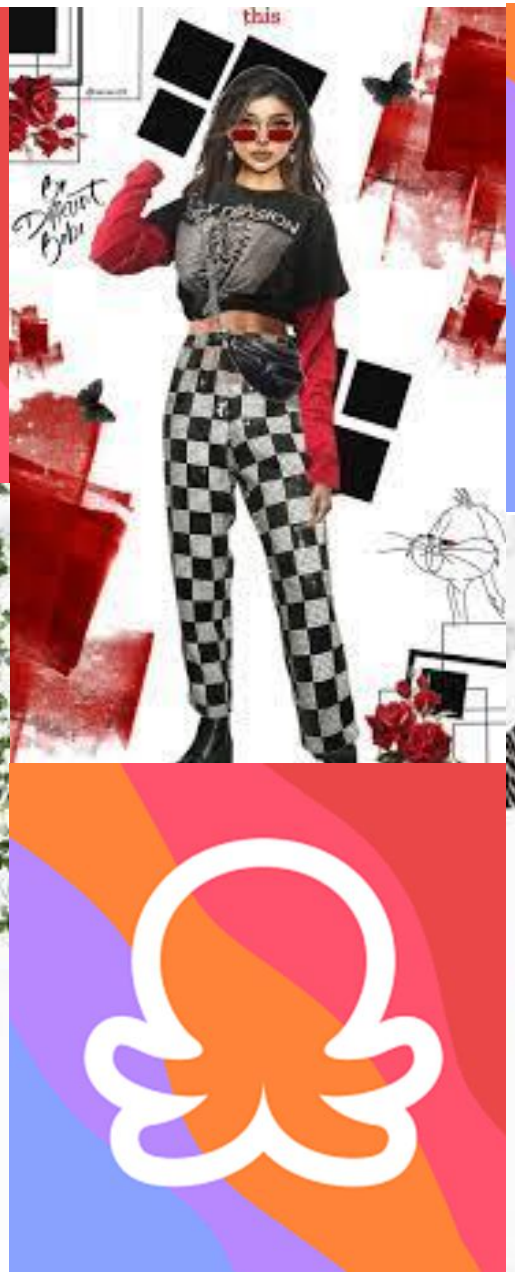

Combyne

Data Science Case Study

completed by

Seghe Momodu

27/09/2021

1. SQL

Question 1: Calculate the share of new and existing users. Output the month, share of new users, and share of existing users as a ratio.
New users are defined as users who started using services in the current month. Existing users are users who started using services in the current month and used services in any previous month. Assume that the dates are all from the year 2021.

```
/*** all users for the year*****/
WITH Users as(SELECT user_id, time_id,
                FIRST_VALUE(time_id) OVER (PARTITION BY by user_id ORDER BY user_id) 'First_Service',
                LAST_VALUE(time_id) OVER (PARTITION BY user_id ORDER BY user_id) 'Last_Service'
                        FROM TableDS),
/*** separates new and existing customers for the month*****/
                        CustomerType AS (SELECT *,
CASE WHEN First_Service >= DATEADD(mm, DATEDIFF(MM, 0, GETDATE()), 0) THEN 'New Customer’
        ELSE 'Existing Customer’  END AS 'Customer_Type’
        FROM Users
WHERE month(time_id) = month(getdate()))
/*** percentage calculation*****/
SELECT datename(month, time_id) 'Month',    SUM(
CASE WHEN Customer_Type = 'New Customer'
        THEN 1 ELSE 0
END)/CAST(COUNT(*) AS DECIMAL(6,2)) * 100 AS '%New Customers',
SUM(CASE WHEN Customer_Type = 'Existing Customer' THEN 1 ELSE 0 END)/CAST(COUNT(*) AS DECIMAL(6,2)) * 100 AS '%Existing Customers'
FROM CustomerType
GROUP BY datename(month, time_id)
```

2. SQL

Question 2: Find the nth highest followerCount given the followerCount table:

```
DECLARE @nthHighest INT = 2


SELECT TOP(1) * FROM (
        SELECT DISTINCT TOP(@nthHighest) followerCount AS NthHighestFollowerCount
FROM followerCount
ORDER BY followerCount DESC) Highest
ORDER BY NthHighestFollowerCount
```

3. Evaluation of Model Performance

About the data

- The features of events_data were engineered expanded and concatenated with the retention score table. The data was highly sparse with high variance.
- The data was split into train, validation and test sets.
- The Logistic Regression and Random Forest Classifier models have been used to model the test data.
- The LASSO Regression technique was used to select 10 relevant features from the 52 features.
- The Principal Component Analysis (PCA) technique was also used to reduce the dimensionality of the dataset and extract 6 of the most important variables.
- Despite the feature selection and extraction techniques used, the data however maintained its high variance, The table shows the variance of the selected variables

Variable	screen_view	item_added	item_removed	outfit_liked	Item_saved	comment_posted
Variance	244597.017905	12433.100152	2043.037700	2043.037700	1137.952605	97.604778

Model	Accuracy	Precision		Recall		ROC-AUC Score
Target		Churn	Not Churn	Churn	Not Churn	
Logit	0.516179	0.52	0.53	1.00	0.00	0.512994
Random Forest	0.484638	0.50	0.31	0.89	0.05	0.486206

3. Evaluation of Model Performance

About the Models

- The models used underperformed in the classification task, however the Logistic regression model performed relatively better than the Random Forest Model.
- The Logit model achieved an accuracy of 52% approx., where the Random Forest Model had an accuracy of 48% approx.
- Both models returned ROC-AUC scores of 51% (Logit) and 49% (Random Forest). These scores are considerable poor as they it demonstrates a poor ability of the model to differentiate between classes (Not – Churn and Churn).
- The Logit model however was relatively able to a lot of relevant instances, hence a low False Negative Rate.
- The Logit model, however under performed in its predictions as it returned only a few relevant (actually correct) predictions from its retrieved instances. Hence a high False Positive Rate.
- I therefor opine that this models are not adequate for decision making. To improve the efforts of classification models, however, some domain knowledge of the business processes with regards to the user events would help me better choose relevant variable and build models that can perform better at correctly identifying the users that have a high probability to churn.

Model	Accuracy	Precision		Recall		ROC-AUC Score
Target		Churn	Not Churn	Churn	Not Churn	
Logit	0.516179	0.52	0.53	1.00	0.00	0.512994
Random Forest	0.484638	0.50	0.31	0.89	0.05	0.486206

4.Interpretation of Statistical Model Results

Question 4: Build a statistical model explaining churn in terms of the explanatory variables and provide interpretations of coefficients and standard errors in these models.

OLS Regression Results

Dep. Variable:	churn	**R-squared:**	0.763
Model:	OLS	**Adj. R-squared:**	0.763
Method:	Least Squares	**F-statistic:**	3.009e+05
Date:	Mon, 27 Sep 2021	**Prob (F-statistic):**	0.00
Time:	04:16:40	**Log-Likelihood:**	-1794.2
No. Observations:	373296	**AIC:**	3598.
Df Residuals:	373291	**BIC:**	3653.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.2061	0.001	1616.429	0.000	1.205	1.208
screen_view	3.782e-06	7.09e-07	5.337	0.000	2.39e-06	5.17e-06
item_added	-7.713e-05	3.31e-06	-23.287	0.000	-8.36e-05	-7.06e-05
item_removed	3.891e-05	8e-06	4.865	0.000	2.32e-05	5.46e-05
retentionScore	-7.2311	0.007	-1092.793	0.000	-7.244	-7.218

Omnibus:	17079.494	**Durbin-Watson:**	2.000
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	8400.829
Skew:	0.175	**Prob(JB):**	0.00
Kurtosis:	2.354	**Cond. No.**	9.38e+03

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.38e+03. This might indicate that there are strong multicollinearity or other numerical problems.

I have used an OLS (Ordinary Least Squares) regression model to explain churn in terms of the explanatory variables. The figure on the left shows the results for the OLS regression.

The equation for the regression line is:

$$Y = 1.2061 + 0.000003782\ \text{screen\_view}\ (X_1) - 0.00007713\ \text{item\_added}\ (X_2) + 0.00003891\ \text{item\_removed}\ (X_3) - 7.2311\ \text{retentionScore}\ (X_4)$$

Where, Y = Churn (The dependent variable),

X_1 = screen_view,

X_2 = item_added,

X_3 = item_removed

X_4 = retentionScore

4. Interpretation of Results

Interpretation of Coefficients

Churn is expected to increase by **0.000003782** when **screen_view (X_1)** increases by 1, decrease by **0.00007713** when **item_added (X_2)** increases by 1, increase by **0.00003891** when **item_removed (X_3)** increases by 1, decrease by **7.2311** when **retentionScore (X_4)** increases by 1 and is predicted to be **1.2061** (when all the variables are equal to 0.

Interpretation of Standard Error

The coefficients are the standardized effect sizes as they indicate the strength of the relationship between the variables without using the original data units. The standard error gives us an estimate of the standard deviation of the coefficient and the amount by which it varies across cases.

4. Planning

In order to make the feed fully dynamic such items that the user's previous activity and geographically trending outfits are shown:
I shall consider tracking the following:

- Track the posts the user liked
- Track the pages the user viewed
- Track the trending posts in each location
- Track the interests and profiles of the users
- Trending hashtags in the user's location

Another thing to look at is the user's last activity. The following items should be tracked per user

- The user's last activity,
- The user's gender
- The pages followed by the user
- The top-N following of the user
- The top influencers followed by the user
- The locations of the people the user is following
- Trending outfits within the user's location
- Track hashtags from the user's top-N following

When onboarding a new user, I would consider capturing data about the user's preferences.

- The user's personal information such as age, location, gender, sexual orientation,
- The user's preferences regarding biodegradable or eco-friendly products
- The user's lifestyle and culture interests
- The user's favorite brands
- The user's personality and aspirations
- The user's colour and seasonal preferences
- The user's preferred outdoor events such as beach parties, meeting, hiking, picnics, dinner
- The user's fictional and traditional influences such as cosplay, Halloween, Oktoberfest, Mardi Gras, Manga animation

This would help in providing the most relevant feeds for the user.