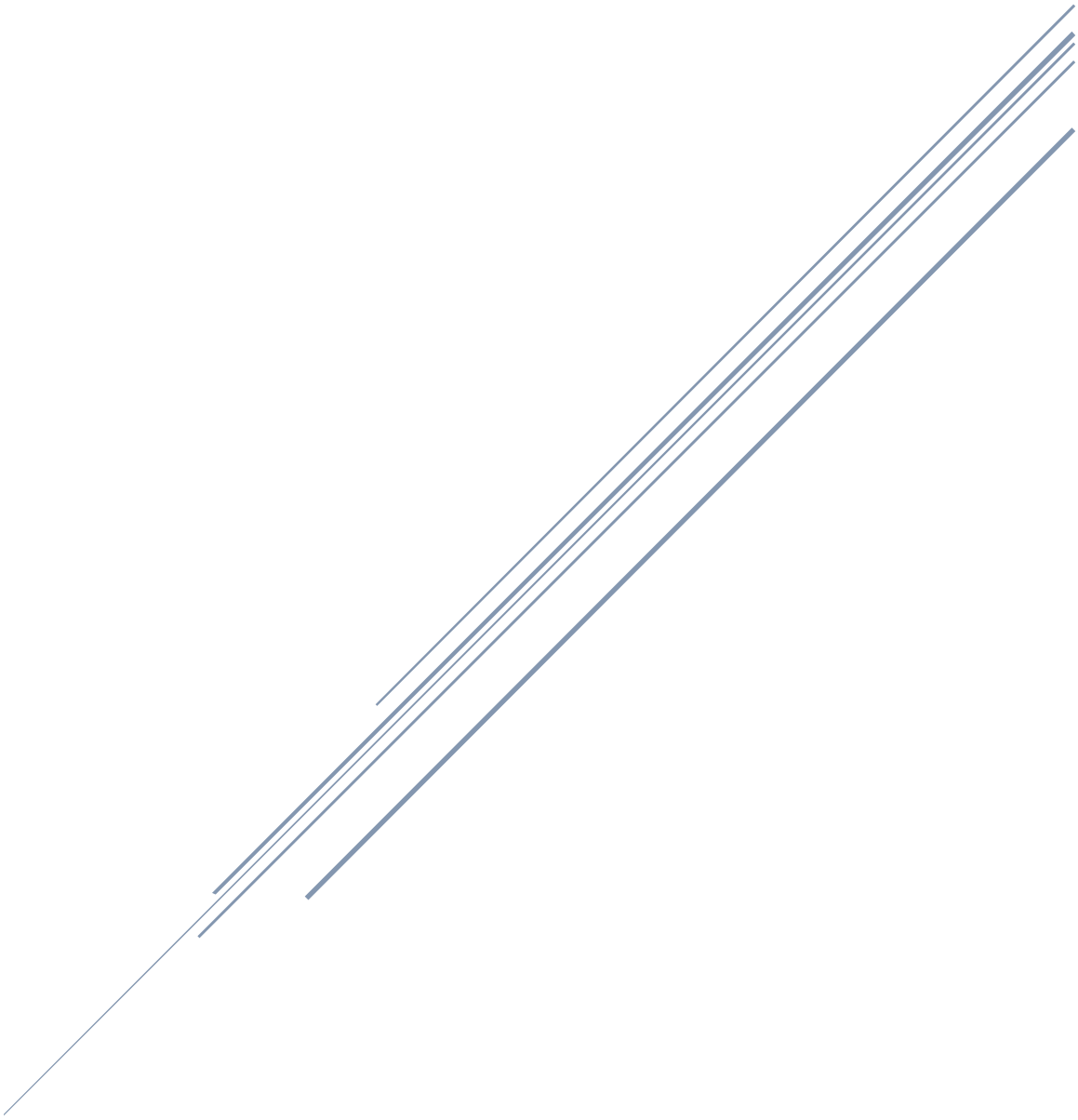
CUSTOMER-TO-CUSTOMER E-COMMERCE BUSINESS PERFORMANCE ANALYSIS REPORT



FEMI BABATUNDE

FEBRUARY 2021

C2C E-COMMERCE BUSSINESS REPORT

Running an online store is different from the conventional physical store. E-commerce business owners take time to ensure the platform is well built to give optimum display of products to consumers, make the platform responsive and easy to navigate, structure the products' pricing in ways that customers are retained and amongst others, ensure that customers are constantly satisfied. These tasks require a lot of skills and it should be noted that investing a lot of money and time into setting up an e-commerce website does not guarantee sales, optimal return-on-investment, and customer retention.

A look at amazon.com, ebay shopify.com and woocomerce show how ecommerce have come to stay in the marketplace. Ecommerce has gradually replaced a lot of brick and mortal businesses and it provides options for both sellers and customers to engage in businesses at their own comfort. Amazon and Ebay are platforms that enable a user to be a buyer and seller on the same platform at the same time. This is an open market model that has come to dominate the ecommerce sphere. On the other hand, shopify and woocomerce (and other hosting platforms) enable business owners to start an ecommerce platform at the comfort of their homes. On the face, the success of any ecommerce platform depends on the number of sales generated by business owners on the platform, activeness of users on the platform and general satisfaction of users (especially buyers).

This project was done to analyze a fashion customer-to-customer (C2C) e-commerce platform that enables users to sell products to other users on the platform. This is a platform similar to Amazon.com but focuses only on sales of fashion related products. As stated earlier, a lot of factors determine the success of a platform, these factors to a large extent impact the favorability

of the platform among sellers (likewise buyers). Performance of a seller on the platform does not only depend on the user interface of the C2C platform, but it is also a function of the type of products uploaded by the seller, the image quality of the product, its description, customer service and social engagement of sellers (The factors are not limited to those stated). Also, buyers' tendencies to purchase products on the platform is a function of some factors like the image of the products, the description of production, established relationship with seller, activeness on the app, amongst other things. In general, while the majority of responsibility of driving sales lies with sellers (sellers need to engage buyers using call to actions), the platform also needs to provide an easy-to-use platform for both the buyers and sellers. The C-2-C platform under review has an embedded social medial platform that enables users (buyers and sellers) to follow each other. This is assumed to enable sellers to build relationship with the buyers. Another assumption is that the social media integration on the platform provides structure through which products made available by sellers can only be viewed by other users that follow them.

The data from the C-2-C platform was sourced from Kaggle.com. It consists of 98,000 observations and 24 fields. The following objectives were set out to be achieved using the available dataset:

- Identify the factors that contribute to sellers being able to generate good sales from the e-commerce platform
- Establish the typical lifetime value of a customer on the platform?
- Establish the average retention rate of buyers on the e-commerce platform?
- Considering that the platform is situated in France, what is the tendency that other users from other countries will sign up on the platform.

- How active are users generally on the e-commerce platform?

DATA SCIENCE APPROACH TAKEN:

DATA WRANGLING

Data collected from Kaggle was made up of 98,000 observations and 24 characteristics originally. Considering the nature of the data, not much wrangling was done on the dataset. It is worth noting that the original dataset was in French (the dataset belongs to a French e-commerce platform), attempt was made to interpret the dataset to English. Due to the long time required for the interpretation, frequently occurring words in the dataset were interpreted to identify. Also, the characteristics captured in French were eventually processed to dummy variables so that they could be used during modelling stage of the project. A total of 229 features were engineered from the original 24 characteristics present in the original dataset.

Fields	Description
identifierHash	Hash of User ID
type	The entity type
country	User's Country (written in French)
language	The User's Preferred language
socialNBFollowers	Number of users who subscribed to this user's ...
socialNbFollows	Number of user account this user follows. New ...
socialProductsLiked	Number of products this user liked
productsListed	Number of currently unsold products that this ...
productsSold	Number of products this user has sold
productsPassRate	% of products meeting the product description....
productsWished	Number of products this user added to his/her ...
productsBought	Number of products this user bought
gender	user's gender
civilityGenderId	civility as integer
civilityTitle	Civility Title
hasAnyApp	user has ever used any of the store's official...
hasAndroidApp	user has ever used the official Android app
hasIosApp	user has ever used the official iOS app
hasProfilePicture	user has a custom profile picture
daysSinceLastLogin	Number of days since the last login
seniority	Number of days since the user registered
seniorityAsMonths	See seniority in months
seniorityAsYears	See seniority in years
countryCode	user's country (ISO-3166-1)

Table 1: Original Dataset Fields

DATA EXPLORATION

The dataset was divided along the line of user's activities on the platform. A user is considered a seller if such user has sold at least one product on the platform. On the other hand, a user is considered a buyer if at least 1 product has been purchased on the platform. 2% of the dataset fell under sellers' segment and 5% of the dataset fell under buyer' segment. A total of 12,027 products were captured sold on the platform while 17,006 products were bought on the platform. It was expected that products sold, and products bought were supposed to be equal. An explanation for this difference is that 1% of the sellers also bought products on the platform (the question that remains is that the products were bought from whom?). Hence, an assumption that follows this is that sellers could record a fulfill order on the platform, which will display as bought without having financial transaction taking place on such account (To have a good explanation for this difference, the method of count for both products bought and sold on the platform needs to be examined). 93% of the dataset are users that did not participate in any form of transaction prior to the time the data was pulled on the platform.

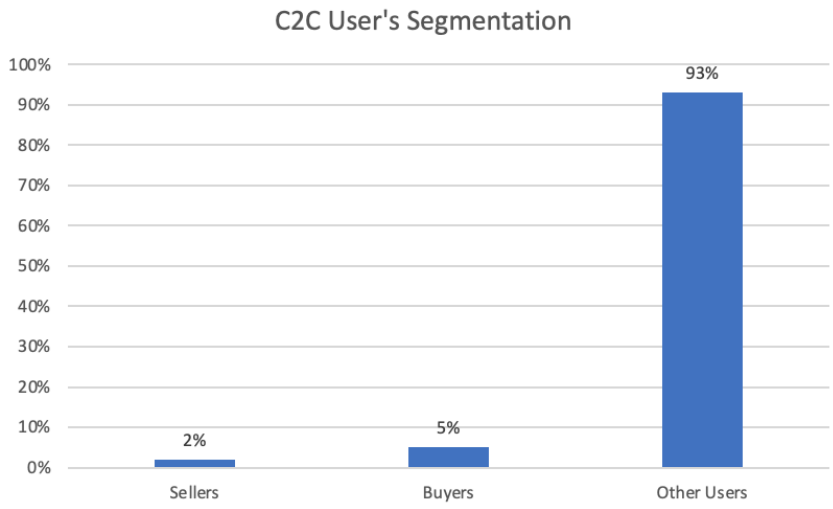


Figure 1: C2C User's Segmentation

It is understandable that a lot of individuals sign up on different platforms without really participating in any form of business transaction, and good number of these platforms mandate

users to create account before being able to access the features of the web page/app. Hence, this results in high number of redundant users on the platform. An evidence of which we can see on this C2C platform.

Sellers' Exploration

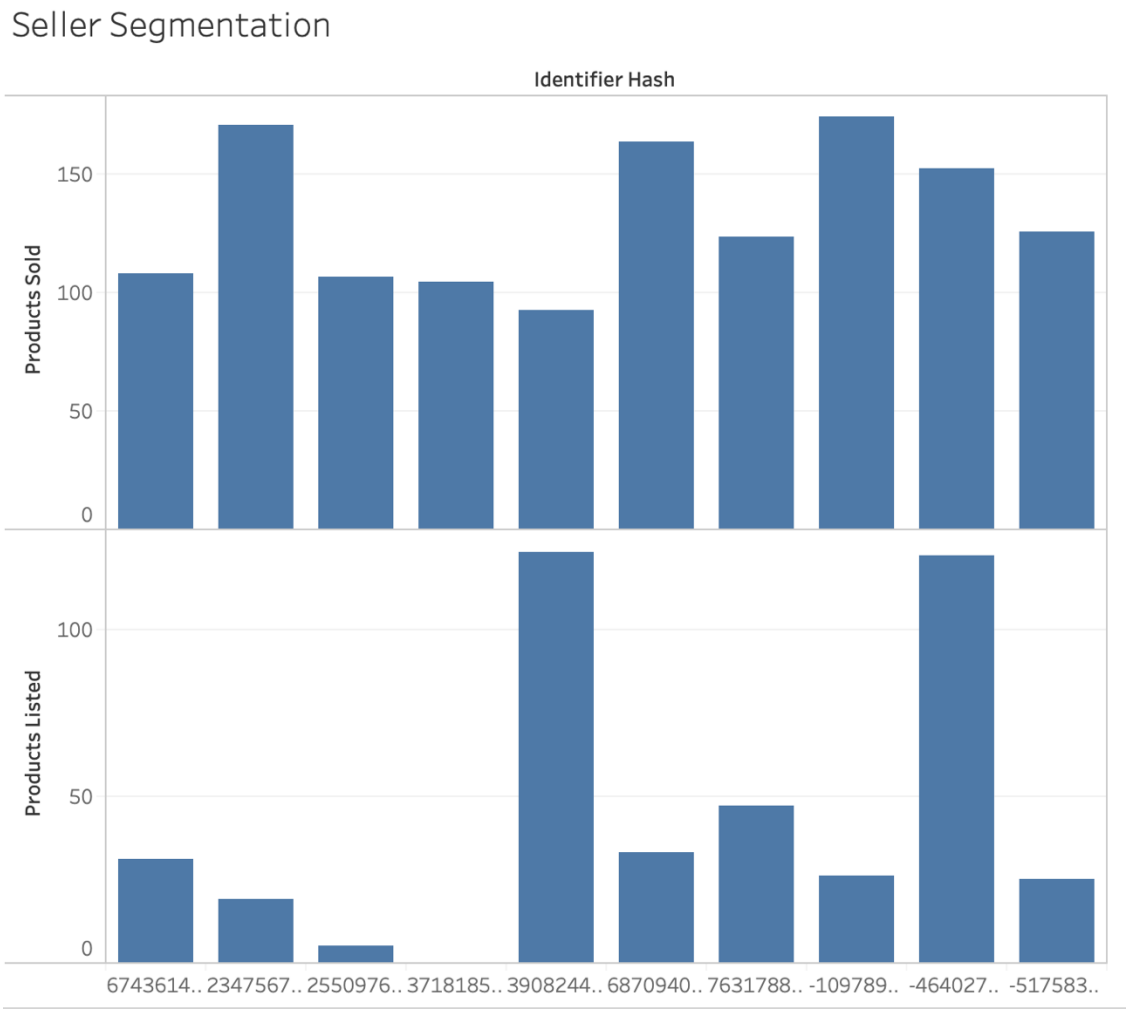


Figure 2: Seller's Products Listed and Products Sold

9 of the top 10 sellers on the platform had products listed for purchase as at the time the data was pulled. Most product sold by any user on the platform was 174.

Using figure 3 below, most products were sold by French users on the platform. Italy, the United Kingdom, Spain, and the United States form the top 5 countries of users where products were

sold. Furthermore, there were only 42 countries of the world where at least one product was sold on the platform.

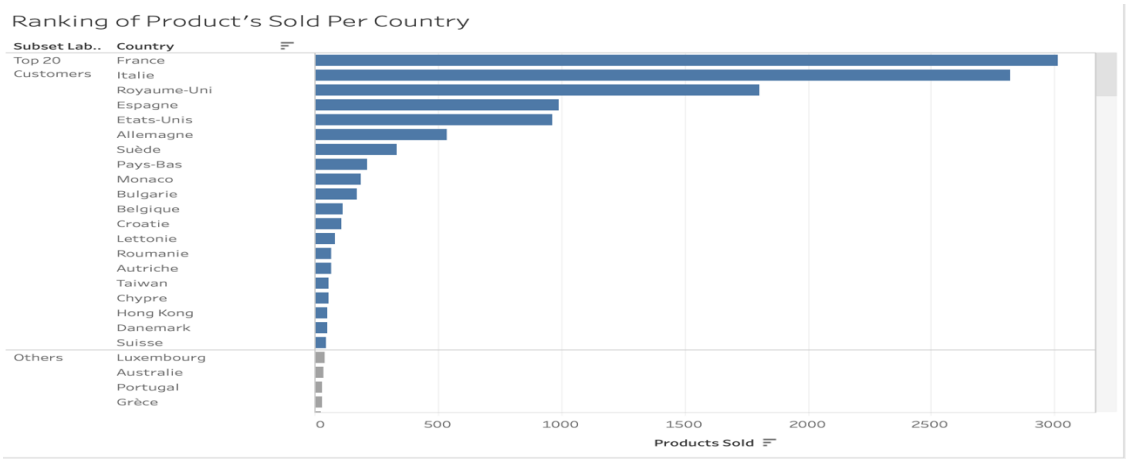


Figure 3: Ranking of Products Sold Per Country for top 20 countries

Considering the platform as that with integrated social media platform, it is obvious that good number of the sellers on the platform have strong followership. The user with the highest number of followers (744), sold 104 products on the platform. Also, the user that sold most on the platform had only 147 followers. As it will be seen later on the heatmap, there is a strong correlation between the number of products sold and followers on the platform. On the flipside, using the heatmap, the number of persons sellers follow have a very weak positive relationship with product sold. Figure 4 captures the distribution of products sold with respect to number of followers on the platform.

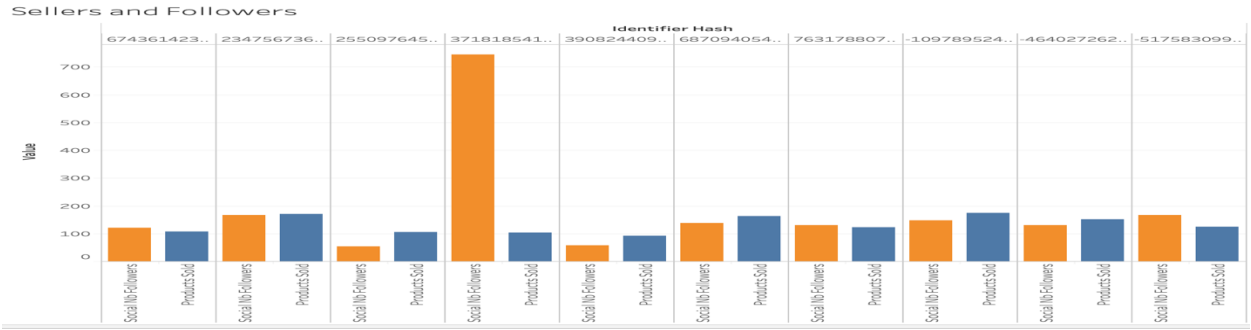


Figure 4: Product Sold and Number of Followers

Sellers' recency on the platform shows that top sellers were more recent on the platform in general. Also, assumption was made that the data was pulled 11 days after it was collected as the least days since last login was 11 days. Using broadcast method, 11 days was subtracted from all of the observations on the platform using the days since last login column. The heatmap also showed that there is a weak negative relationship between days since last login and the number of products sold. Intuitively, this was expected as more active sellers on the platform stand the chance on getting more sales. On average, sellers average last login days was 180 days.

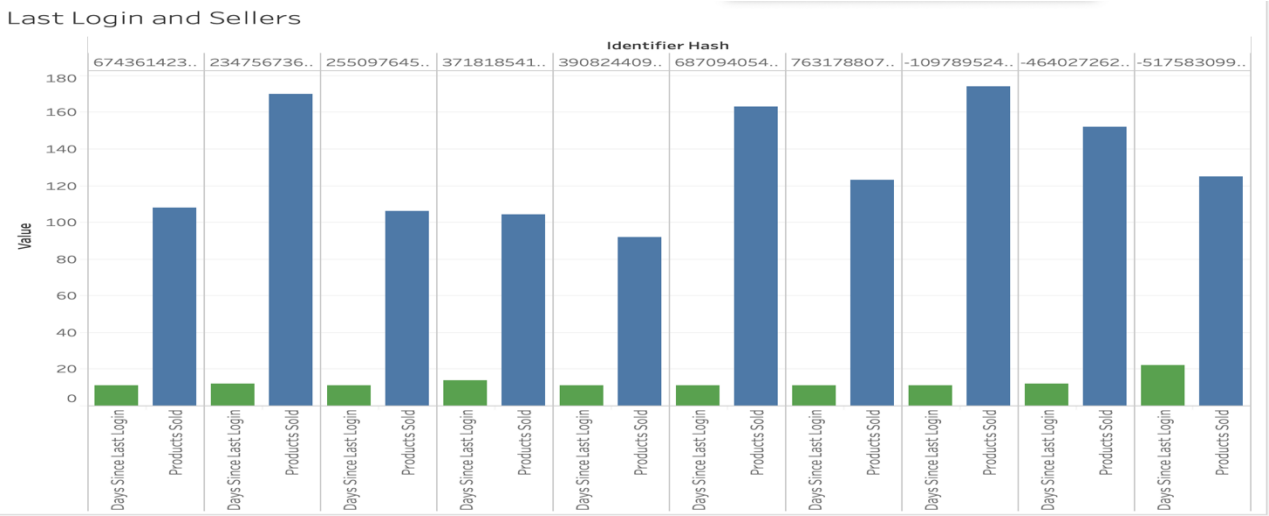


Figure 5: Product Sold and Last Login for top 10 sellers

Buyers' Exploration

As explained above, 5% of the dataset bought atleast one product on the platform. As depicted in figure 6, the most product bought by a user was 405 and this user also had a product listed on the platform. There were 562 users that had participated in both buying and selling of at least one product on the platform. It can also be seen that some of the top buyers on the platform did not sell or list any product on it.

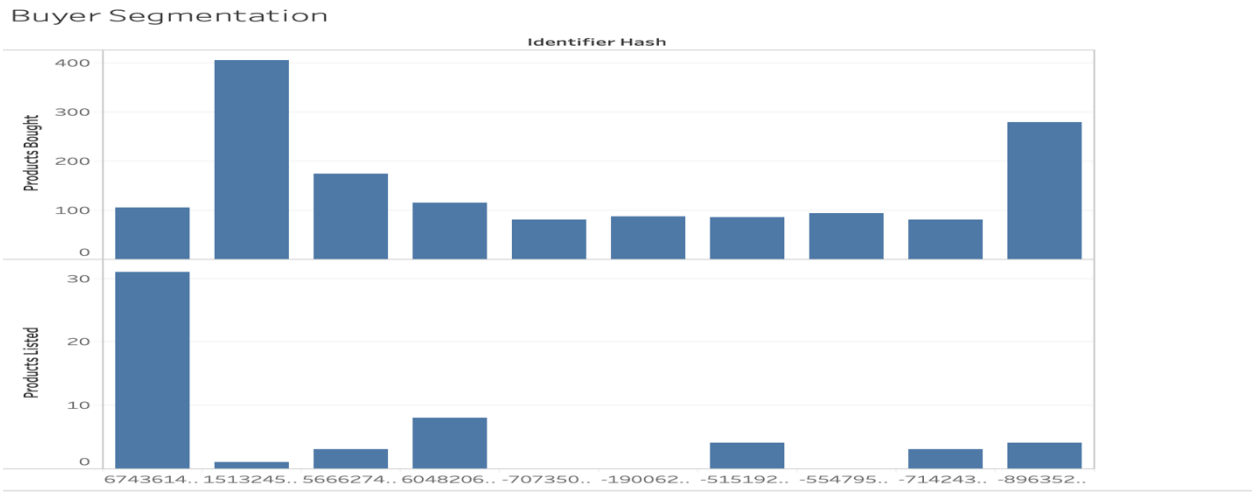


Figure 6: Products Bought and Products Listed

The most products purchased were from France. USA, UK, Germany and Italy were the remaining countries in the top 5 of purchases made by country. Compared to 42 countries that had products sold registered, only 35 countries had users made at least one purchases on the platform.

country	productsBought
France	3573
Etats-Unis	2370
Royaume-Uni	2174
Allemagne	1635
Italie	1221
Espagne	1028
Belgique	718
Suède	566
Pays-Bas	537
Danemark	438

As it can be seen on figure 7, top 10 purchasers on the C2C platform had at least 6 users following them on the platform. On the other hand, there was at least one user in the top 10 that did not follow any other user. The user with most purchases followed only 8 other users. The relationship between products bought and each of number of followers and number of follows can be examined better using the heatmap.

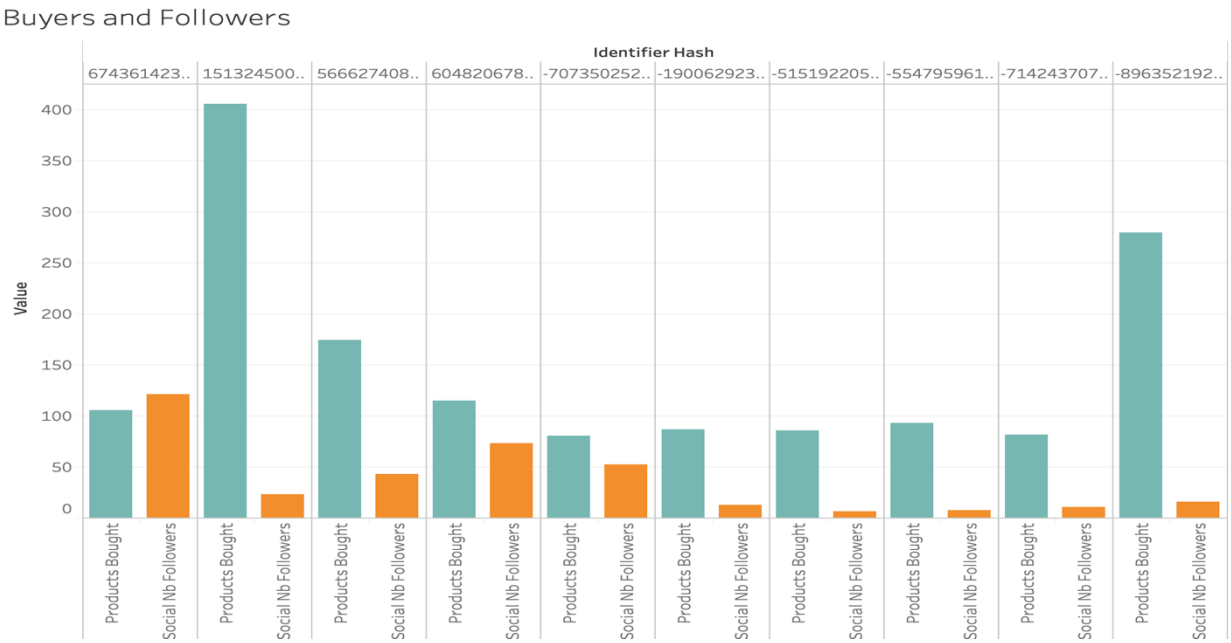
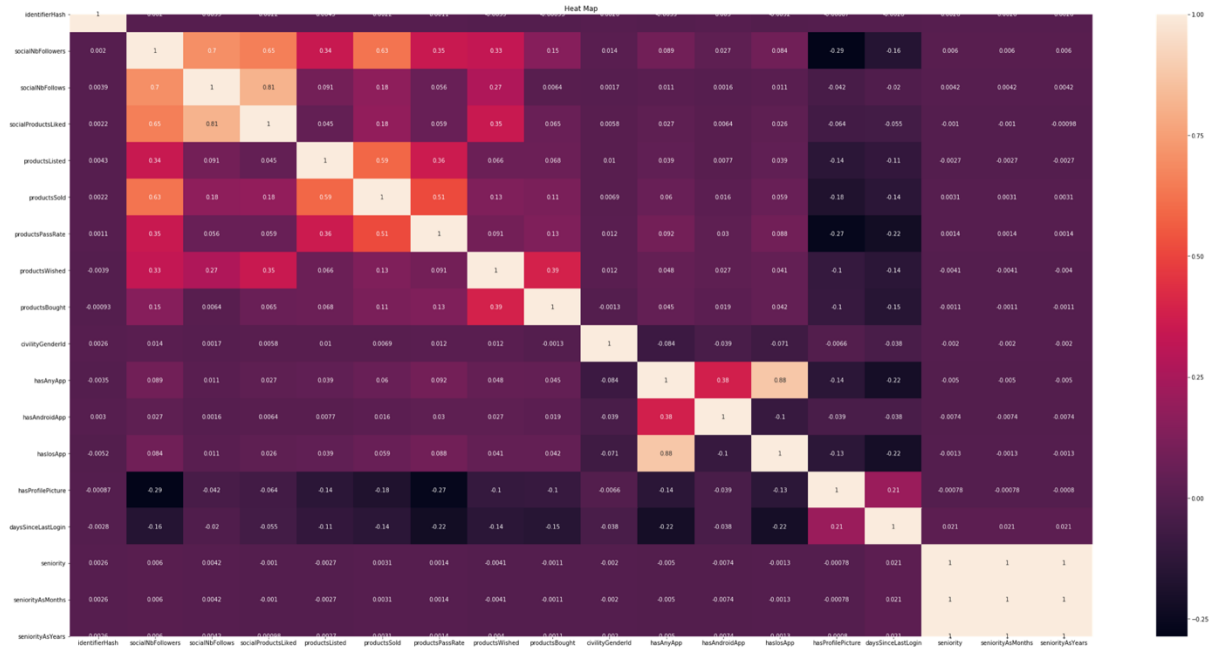


Figure 7: Number of Products Bought and Social Followers

Going by the 11 days assumption made under the sellers' exploratory analysis above, 8 of the top 10 buyers were on the platform on the day the data was collected. However, the average days of last login by users categorized as buyers was 279 (this is roughly 9 months). For users that engaged in no single transaction on the platform, the average number of days since last login was 593 (this is roughly 1 year and 7 months).



From the heatmap above, there are obvious strong relationships between some of the features in the dataframe.

The focus of this correlation matrix will be on productsSold and productsBought. ProductsSold had strong relationship with SocialNbfollowers, productsListed, and productsPassRate in this decreasing order. Weak positive relationships exist with socialNBFollows, socialProductsLiked, productsWished and productsBought. The strong correlation relationships seen between productsSold and some of the variables are expected to exist. The not so very strong relationship with productsBought is a bit surprising. Weak negative relationships exist between productsSold and hasProfilePicture. Also, negative relationship is also seen with daysSinceLastLogin. The negative relationship shows that the more a seller is away from the platform, the lesser he sells. This is intuitive. The only thing here is that it is a weak negative relationship.

On the other hand, as expected, productsBought got the strongest correlation with productsWished. productsBought showed weak prostive relationships with productspassRate and socialNbFollowers. The same negative relationships seen with productsSold exist with

productsBought. The relationship between productsListed and productsBought was positive but very weak (in fact, it can be assumed there is no relationship between these variables).

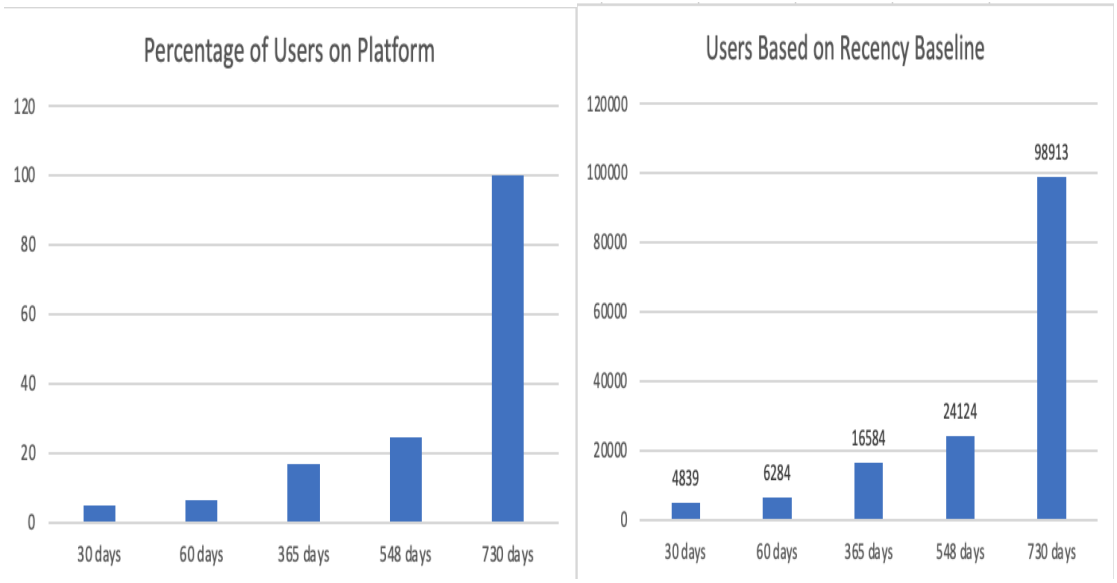
On a general note, productsWished showed strong positive relationships with social activeness on the platform. Is it safe to say a lot of people go on the platform to make social connections or "windowshop"?

The strong positive relationship between number of followers and number followed by users was expected. Also, products liked strongly correlated with SocialNbfollowers and SocialNbfollows. socialProductsLiked had a weak positive correlation with productsSold, productsWished and productsBought (the relationship was positive but very weak as stated above). The relationship between productsListed and productsBought was positive but very weak. PassRate and socialNBFollowers were both positively correlated to productListed.

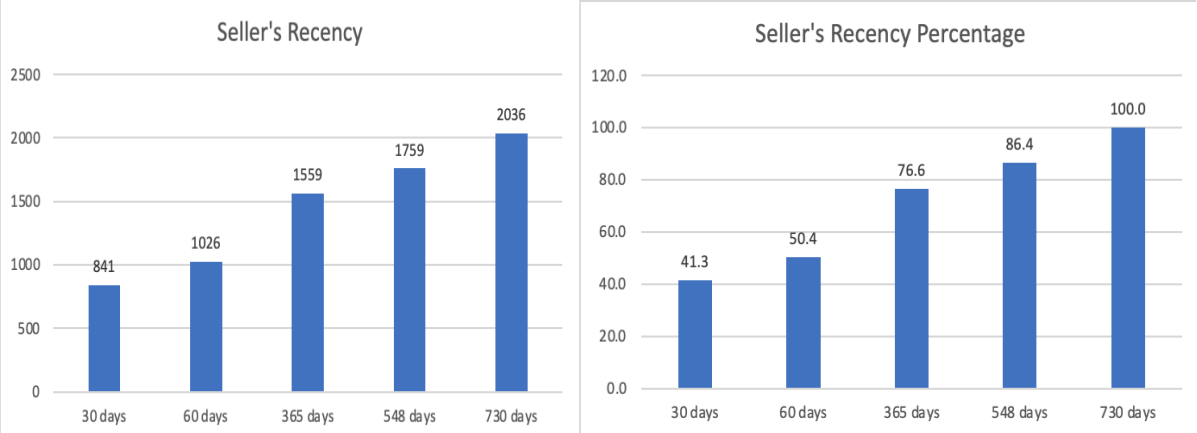
There was multicollinearity between seniority, seniorityAsMonths and seniorityAsYears. This was expected because seniorityAsMonths and seniorityAsYears were derivatives of seniority. For modelling purpose, seniorityAsMonths and seniorityAsYears were dropped off and the seniority feature alone was used along other features.

USERS' ACTIVENESS

Since different organizations have different baseline for retention rate, for this study, days since last login will be used to proxy retention. Users' days since last login was compared for 30 days, 6 months, 1 year, and 2 years and the following results were gotten.

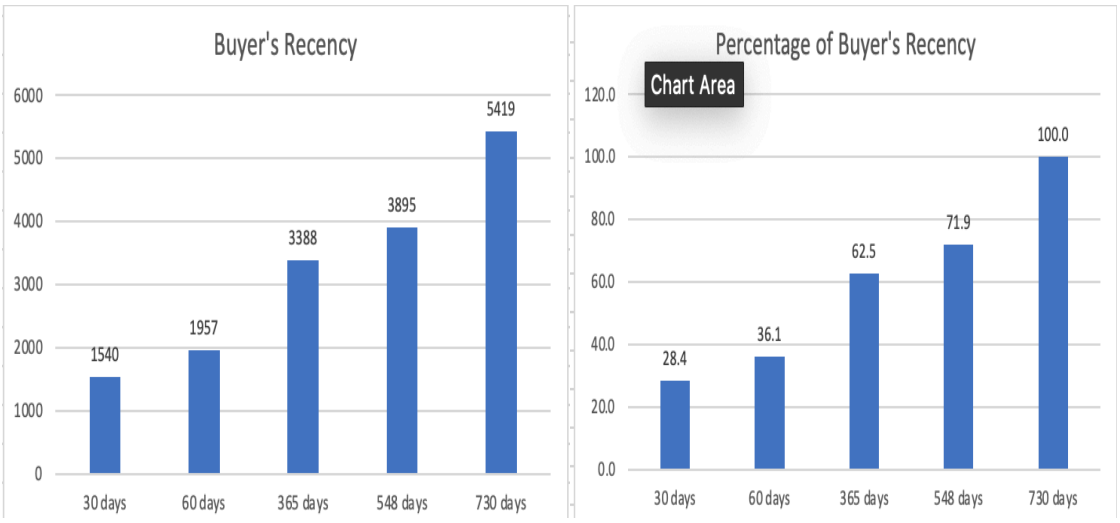


From the table above, it is obvious that approximately 5% of the total users were on the platform within 30 days before the data was pulled. A year just before the data was pulled, 17% (16,584) of the users were active on the platform.



Seller's Recency Plot

Going by the seller's recency plot, it can be seen that 41.3% of the sellers were active on the platform within the first 30 days before the data was pulled. More than 70% of the users considered as sellers were also active within the first 1 year before the data was pulled.

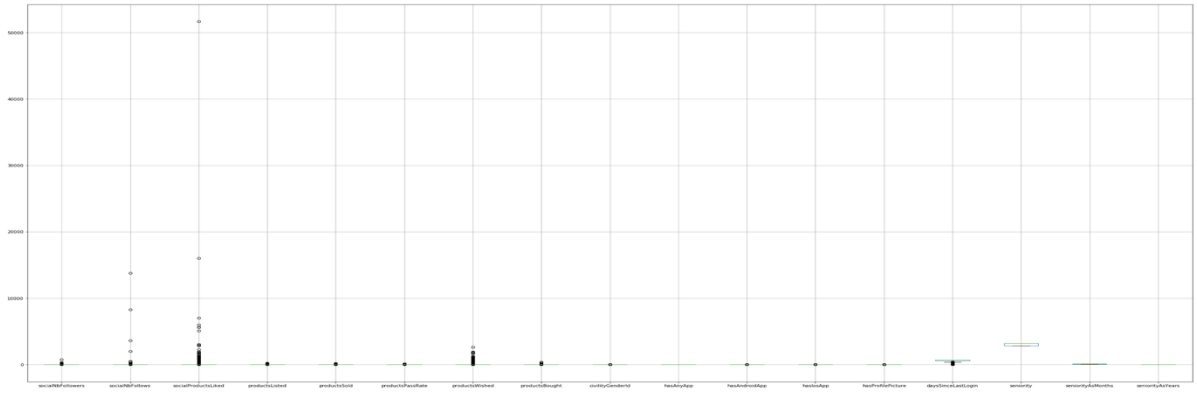


Buyer's Recency Plot

The buyer's recency plot shows that 28.4% of the users of the buyers were on the platform just 30 days before the data was pulled. In a year, 62.5% of the buyers were active on the platform 1 year before the data was pulled. It is obvious that sellers were more active than buyers.

DEALING WITH OUTLIERS

Outliers were noticed on the dataset, but due to the sizes of buyers and sellers being 2% and 5% of the data frame respectively, decision was made not to drop these outliers. Dropping the outliers would have drastically affected important observations in the dataset and this will definitely affect the final result that will be gotten from the model.



Boxplot to show outliers

DATA PROCESSING

The data processing step involved the normalization of the dataset as good number of the variables were skewed to the right. The normalization was equally necessary as it was required for standardization of the data frame.

```
col_names = ['socialNbFollowers',
       'socialNbFollows', 'socialProductsLiked', 'productsListed',
       'productsSold', 'productsPassRate', 'productsWished', 'productsBought',
       'daysSinceLastLogin',
       'seniority']
dflog = np.log(dfProcessed[col_names]+1)
dflog.describe()
```

```
from sklearn.preprocessing import StandardScaler

SS_scaler = StandardScaler()

col_names = ['socialNbFollowers',
       'socialNbFollows', 'socialProductsLiked', 'productsListed',
       'productsSold', 'productsPassRate', 'productsWished', 'productsBought',
       'daysSinceLastLogin',
       'seniority']

scaledCols = ['ScSocialNbFollowers',
       'ScsocialNbFollows', 'ScsocialProductsLiked', 'ScproductsListed',
       'ScproductsSold', 'ScproductsPassRate', 'ScproductsWished', 'ScproductsBought',
       'ScdaysSinceLastLogin',
       'Scseniority']

dfScale = SS_scaler.fit_transform(dflog[col_names])

dfScale = pd.DataFrame(dfScale , columns=scaledCols)



dfScale.head()
```

Using the productSold feature as label (for the seller's dataset), the dataset was split into train and test dataset. 25% of the scaled dataset reserved was reserved as the test dataset. Also, the productsBought feature was used as a label for the split buyer's dataset. 25% of the buyer's dataset was also retained as test dataset.

DATA MODELLING

Using sklearn's feature_selection, SelectKBest, 10 most important features were selected for the dummy regressor. Feature selection is a process where you automatically select those features in

your data that contribute most to the prediction variable or output in which you are interested. Having irrelevant features in your data can decrease the accuracy of many models, especially linear algorithms like linear and logistic regression.

Three benefits of performing feature selection before modeling your data are:

- **Reduces Overfitting**: Less redundant data means less opportunity to make decisions based on noise.
- **Improves Accuracy**: Less misleading data means modeling accuracy improves.
- **Reduces Training Time**: Less data means that algorithms train faster.[1]

The score_function used with the SelectKBest was f_regression which uses F-value between label/feature for regression tasks[2].

Seller's Models

Using the dummy regressor, the R-squared score on the train dataset was 0.0 and for the test dataset the score was equally 0.0 (real value was -0.000132). This served as the baseline for other models used for the study.

Using multiple linear regression, the MAE was 0.113, the train set R-square was 81.6% and the test score was 79.4%. Using KFold of 5 cross-validation, the mean R-squared score was 81%. This confirms the validity of the linear regression model on the seller's dataset. Using grid search with linear regression model, the train scores and the test scores for different R-squared of varied number of features are depicted in the graph below. It is obvious that the model worked well with predicting the dataset.

[1] https://machinelearningmastery.com/feature-selection-machine-learning-python/
[2] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

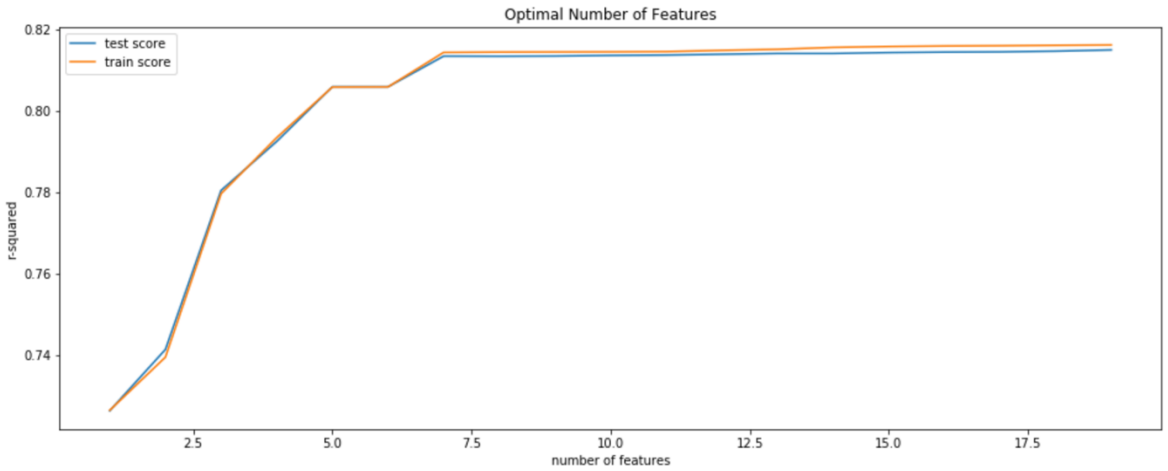


Figure: GridSearch Rsquared output with LR Model

Lasso regression was also used with the seller's dataset. The R-squared gotten for both the train and test datasets before using any cross-validation were 51% and 53% respectively. This result was gotten with alpha at 0.5. Using varied lambda values with cross_val_score at K= 5 while lambda was varied from 0.001 to 0.5 with 20 equal spaced values, the best lambda score was 0.053 and the test R-squared value was 79.73%. Using the best alpha value with the Lasso model on the data set, the train score was 81% and the same test score of 79.7% was gotten.

USING ENSEMBLE METHOD

Boosting as we know is a sequential process, where each subsequent model attempts to correct the errors of the previous model. The succeeding models are dependent on the previous model. Let's understand the way boosting works in the below steps.

1. A subset is created from the original dataset.
2. Initially, all data points are given equal weights.
3. A base model is created on this subset.
4. This model is used to make predictions on the whole dataset.
5. Errors are calculated using the actual values and predicted values.

6. The observations which are incorrectly predicted, are given higher weights. (Here, the three misclassified blue-plus points will be given higher weights)
7. Another model is created, and predictions are made on the dataset. (This model tries to correct the errors from the previous model)

Thus, the boosting algorithm combines a number of weak learners to form a strong learner. The individual models would not perform well on the entire dataset, but they work well for some part of the dataset. Thus, each model actually boosts the performance of the ensemble.3

To view if there will be any improvement seen on the model using boosting method, three ensemble models were used with the dataset, Adaboost, GradientBoost and XGBoost.

Adaboost performed very bad on the dataset as the R-squared scores were in the negatives. On the other hand, GradientBoost's cross-validation score for train dataset gave a result of 85.67%. Using the GradientBoost model on the seller's dataset, train score was 90.16% and the test score was 83.47%. This is a big improvement on the Linear regression model and Lasso regression model. XGBoost's performance was not as good as GradientBoost's and LinearRegression models. The table below compare the R-square scores for the four models used for the prediction.

	Algorithm	RSquare train score	RSquare test score
0	Linear Regression	81.608258	79.428867
1	Lasso Regression	81.073932	79.734204
2	Gradient Boost	90.163292	83.465089
3	XGBOOST	71.908976	71.070795

Table: Comparing the train and test scores for 4 models

[3] https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/

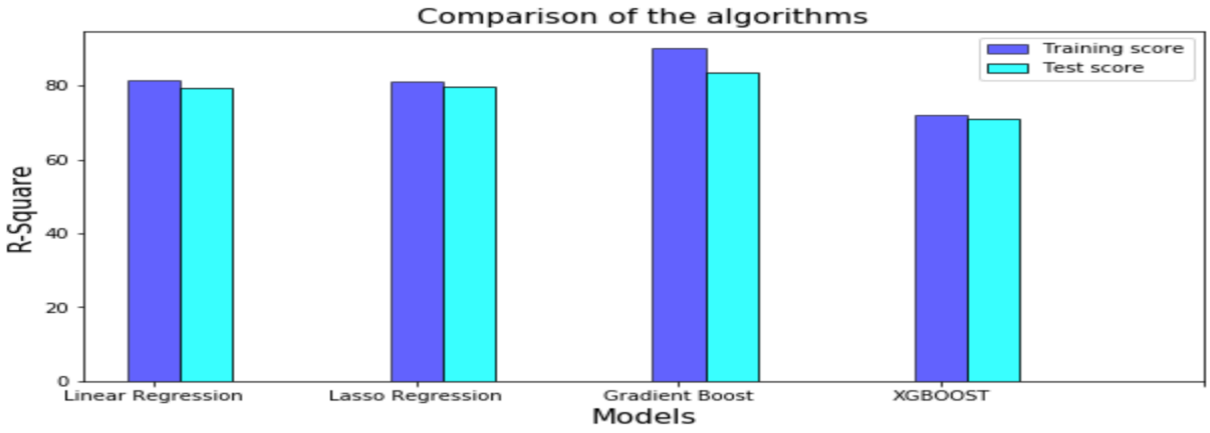


Figure: Model comparison

Going by the table and figure above comparing the models, GradientBoost was the best performing model.

	Coefficient
ScproductsPassRate	0.660306
ScproductsListed	0.189246
ScSocialNbFollowers	0.101497
ScdaysSinceLastLogin	0.016673
hasProfilePicture	0.006528
ScsocialProductsLiked	0.006478
ScsocialNbFollows	0.004969
ScproductsBought	0.003743
ScproductsWished	0.003350
civilityTitle_miss	0.001706

Table: 10 important features of GradientBoost Model

The ten most important features using GradientBoost model can be seen in the table above. Based on these standardized coefficients, it can be seen that a seller has a tendency to make a sale on the platform if the seller's products have high pass rates. Also, it is intuitive to note that the more sellers list products on the platform, the higher chances they have in making sales on the platform. Activeness in terms of being frequent on the platform and socially engaging (uploading profile picture, following other users and increasing followers) with other users will improve the ability of sellers to make more sales on the platform. Since this platform is a socially structure one, there

is tendency for a seller to generate more sales if such seller patronizes other sellers or at least shows interest in other sellers' products. Single females have higher chance of being sellers on the platform.

Hypothesis testing was done on three of the top features (productPassRate, productListed and nbSocialFollowers) for this model and the three features were significant at 1% level of significance.

Buyer's Model

The same approach used for the seller's dataset was used for the buyer's dataset. A dummy regressor was used on the buyer's dataset, the mean absolute error of the model was 0.392 and R-squared for both train and test datasets were 0.0 (test set score was more into negative but close to 0.0).

Using multiple linear regression, the MAE was 0.317, the train set R-square was 32.3% and the test score was 34.4%. Using KFold of 5 cross-validation, the mean R-squared score was 32%. This confirms the validity of the linear regression model on the seller's dataset. Using grid search with linear regression model, the train scores and the test scores for different R-squared of varied number of features are depicted in the graph below. The R-Squared is quite low compared to what was experienced with the seller's dataset. This shows that the regressors are not good predictors of the dependent variables.

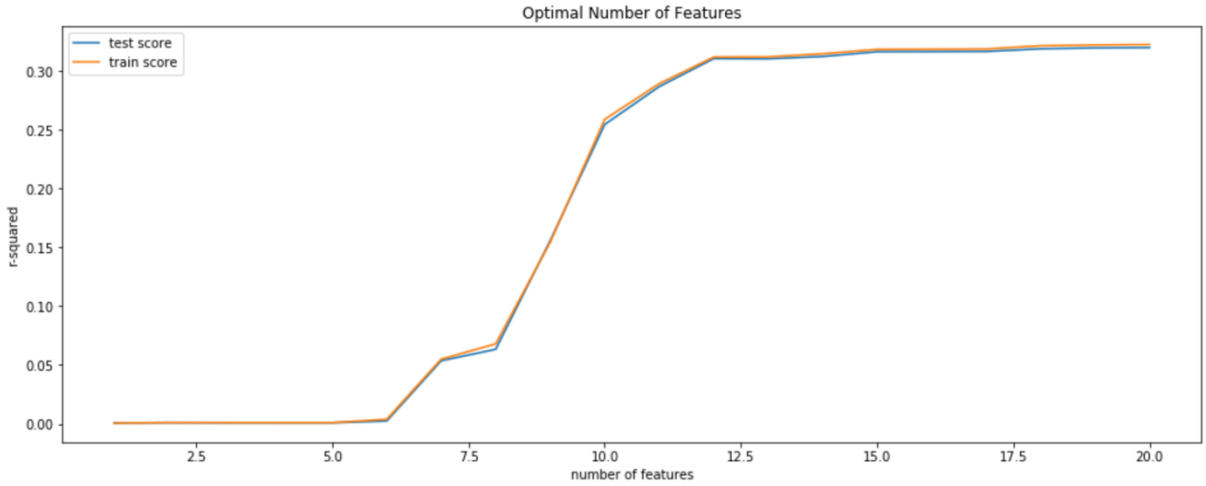


Figure: GridSearch R-squared output with LR Model on Buyer's Data set

Lasso regression was also used with the buyer's dataset. The R-squared gotten for both the train and test datasets before using any cross-validation were 31% and 34% respectively. This result is very similar to what was gotten with the multiple linear regression model. This result was gotten with alpha at 0.5. Using varied lambda values with cross_val_score at K= 5 while lambda was varied from 0.001 to 0.5 with 20 equal spaced values, the best lambda score was 0.001 and the test R-squared value was 34.4%. Using the best alpha value with the Lasso model on the data set, the train score was 32.2% and the same test score of 34.4% was gotten.

USING ENSEMBLE MODEL

To view if there will be any improvement seen on the model using boosting method, three ensemble models were used with the dataset, Adaboost, GradientBoost and XGBoost.

Adaboost performed very bad on the buyer's dataset as the R-squared scores were in the negatives. On the other hand, GradientBoost's cross-validation score for train dataset gave a result of 38.6%. Using the GradientBoost model on the seller's dataset, train score was 45% and the test score was 41%. This is a big improvement on the Linear regression model and Lasso regression model.

XGBoost's performance was not as good as GradientBoost's and LinearRegression models. The table below compare the R-square scores for the four models used for the prediction.

	Algorithm	RSquare train score	RSquare test score
0	Linear Regression	32.260299	34.410983
1	Lasso Regression	32.243275	34.402929
2	Gradient Boost	45.038247	41.593037
3	XGBOOST	28.265453	29.279595

Table: Comparing the train and test scores for 4 models

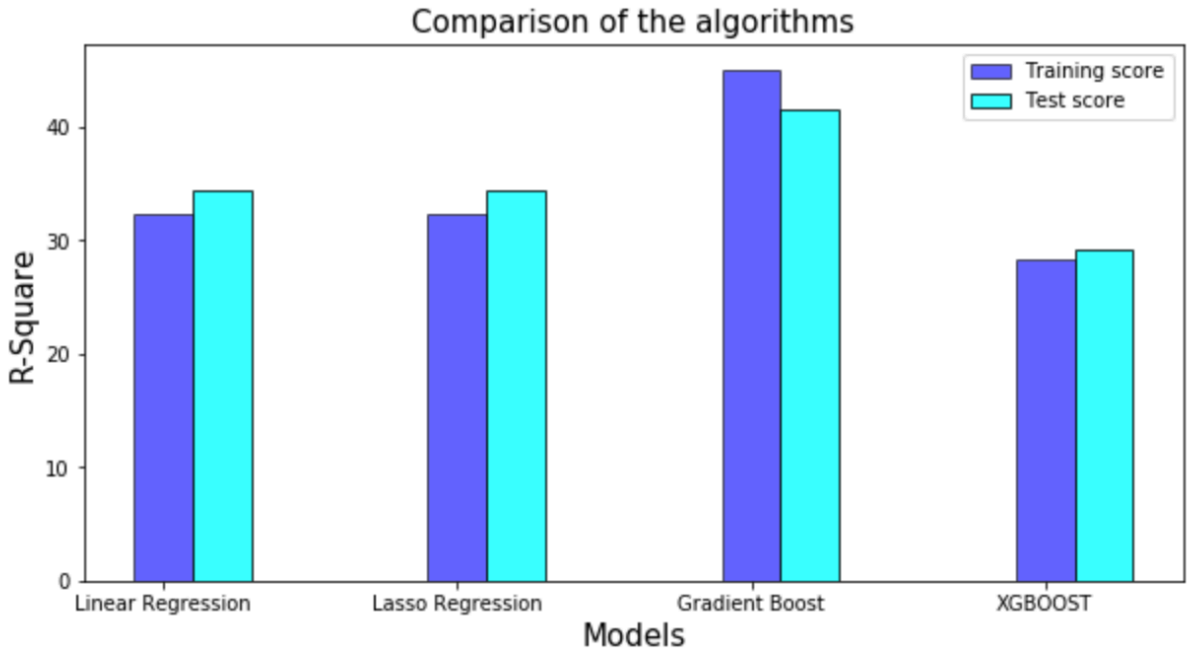


Figure: Model comparison

Going by the table and figure above comparing the models, GradientBoost was the best performing model.

	Coefficient
ScproductsWished	0.157994
ScdaysSinceLastLogin	0.116070
ScsocialProductsLiked	0.106573
ScSocialNbFollowers	0.101904
ScsocialNbFollows	0.036696
hasProfilePicture	0.015952
gender_F	0.015249
civilityTitle_mrs	0.014962
ScproductsSold	0.010446
ScproductsPassRate	0.008960

The table above shows the best performing features when GB was used with the buyer's dataset. As expected, if a buyer indicates interest in a product, there is high chance such buyer will eventually buy the product. A user that frequents the platform has high chance of making purchase on the platform. Just as it was seen with the seller's dataset, social activeness of users on the platform will drive users to make purchases on the platform. Buyer's also have high tendency of completing a purchase if the wished product has a good pass rate. As seen while discussing the seller's important features, females generally have more impact on buying and selling on the platform. Ads to products of sellers can be targeted at females to increase sales. On the other hand, social interaction of sellers can be increased with both married and single females on the platform so that they could generate more sales. These feature ranking of the GB model is reasonable as the dataset is for a fashion-based platform and females generally might have more tendencies to patronize and sell on the platform (more study needs to be done to view the interaction of gender with sales on the platform).

Hypothesis testing was done on three of the top features (productsWished, nbSocialFollowers and daysSinceLastLogin) for the model done on buyer's dataset and the three features were significant at 1% level of significance.

IDEAS FOR FUTURE STUDY.

Ecommerce is a large sector that has great chance to benefit from machine learning and good data's availability will go a long way to ensure relevant result and recommendations are given. As seen in this study, important information like when a user signed up on the platform, the total amount spent by user on the platform, cost incurred by users (if any) to use the platform, frequency of purchases, types/categories of products, and number visits to the platform were all not given. If these information were made available, better work on recency would have been done. For business related problems like this, Retention and churn rates of users on the platform can be determine. Also, RFM (recency, frequency and monetary value) of the business as a whole can be estimated. Knowing these values, better decisions can be made by platform owners to determine if users are really interacting well with the platform. Also, for sellers, actions needed to drive sales on the platform can be recommended.

Furthermore, another consideration for future study will be designing good recommendation system targeting buyers based on nature of purchases previously made by such seller or other sellers that purchased similar products. These recommendations can also be given to sellers on the best-selling products listed by other users on the platform.

USEFULNESS OF STUDY

This study will be useful to ecommerce businesses like amazon, ebay, shopify and woocomerce as they can design their platforms in ways to ensure that both buyers and sellers are well satisfied with their shopping experiences on the platform.

GITHUB & TABLEAU LINKS

Notebooks for all the stages of this study can be accessed using the github link below: https://github.com/Femibabz/Capstone_Three

For interaction with tableau for the exploratory data analysis, kindly use the link below: https://public.tableau.com/views/C2CEDAAnalysis/SellersDashboard?:language=en&:display_count=y&:origin=viz_share_link

For comments, discussions and criticisms, kindly send emails with subject C2C business analysis to: ofem.b@gmail.com