Project Documentation

E-commerce Churn

Ayman Adel Wahba

Amira Gomaa Abdelaziz Ibrahim

Osama Tarek Mohamed Ibrahim

Hajer zakaria Emira

Table of contents

Project Requirements & Deliverables

-The project requirements described as follows:

Analyze a dataset for an E-commerce business to investigate possible reasons for churn to possibly implement countermeasures for this issue.

-The project deliverables are as follows:

A presentation to summarize the findings of analysis and recommendations for dealing with churn .

A documentation detailing the steps taken to complete the project.

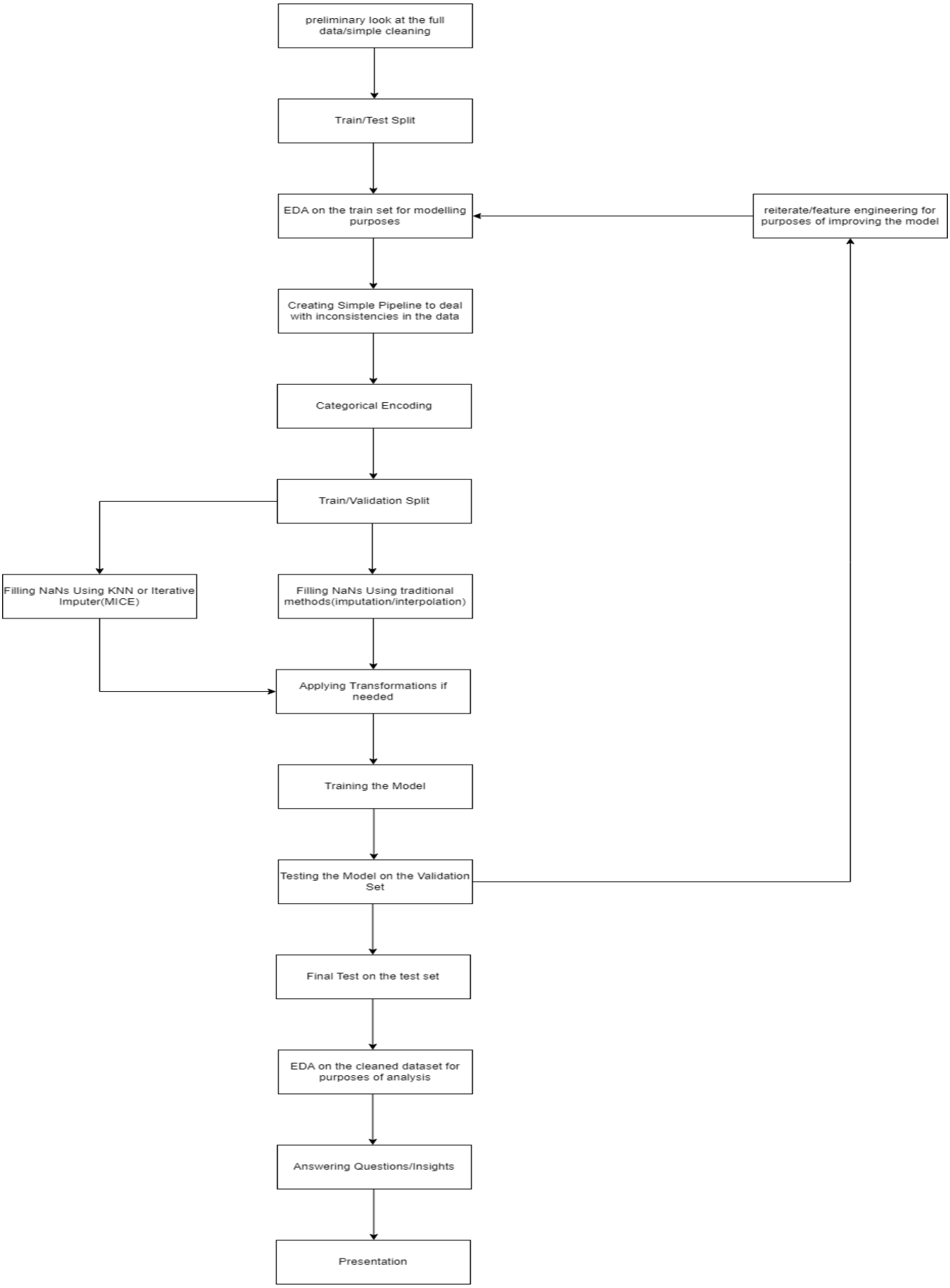
All the code that was written during the project.

Model of Operation

As per the requirements of the project, Analysis and modelling was performed on the dataset, first was the modelling to create a model to predict churning. The reason modelling was done before the EDA was to simulate a real-life scenario where the test data is simply not available to gain a better evaluation of the model that was created. Steps were followed to Clean/Validate and prepare the data for the model which will be described in the project workflow.

-A high Resolution PNG will be provided with the deliverables that details the workflow

-any code that was written to achieve the steps of this model is attached with the documentation.


preliminary look at the full data/simple cleaning
Train/Test Split
EDA on the train set for modelling purposes
reiterate/feature engineering for purposes of improving the model
Creating Simple Pipeline to deal with inconsistencies in the data
Categorical Encoding
Train/Validation Split
Filling NaNs Using KNN or Iterative Imputer(MICE)
Filling NaNs Using traditional methods(imputation/interpolation)
Applying Transformations if needed
Training the Model
Testing the Model on the Validation Set
Final Test on the test set
EDA on the cleaned dataset for purposes of analysis
Answering Questions/Insights
Presentation

Step-1: Modelling

As we have started with modelling, some steps will be performed along with the modelling for the benefit of the whole process and not just exclusively for the modelling such as Duplicate Removal also any steps described moving forward are written and explain thoroughly in the notebooks attached with the documentation.

First, before any splitting was done the data was cleaned of any duplicates and the results of this cleaning was the data shrinking by third of its original size from 16890 down to 5630

CustomerID	Churn	Tenure	PreferredLoginDevice	CityTier	WarehouseToHome	PreferredPaymentMode	Gender	HourSpendOnApp
50510	0	0.0	0	3	9.0	NaN	Male	3.0
50510	0	0.0	0	3	9.0	NaN	Male	3.0
50510	0	0.0	0	3	9.0	NaN	Male	3.0

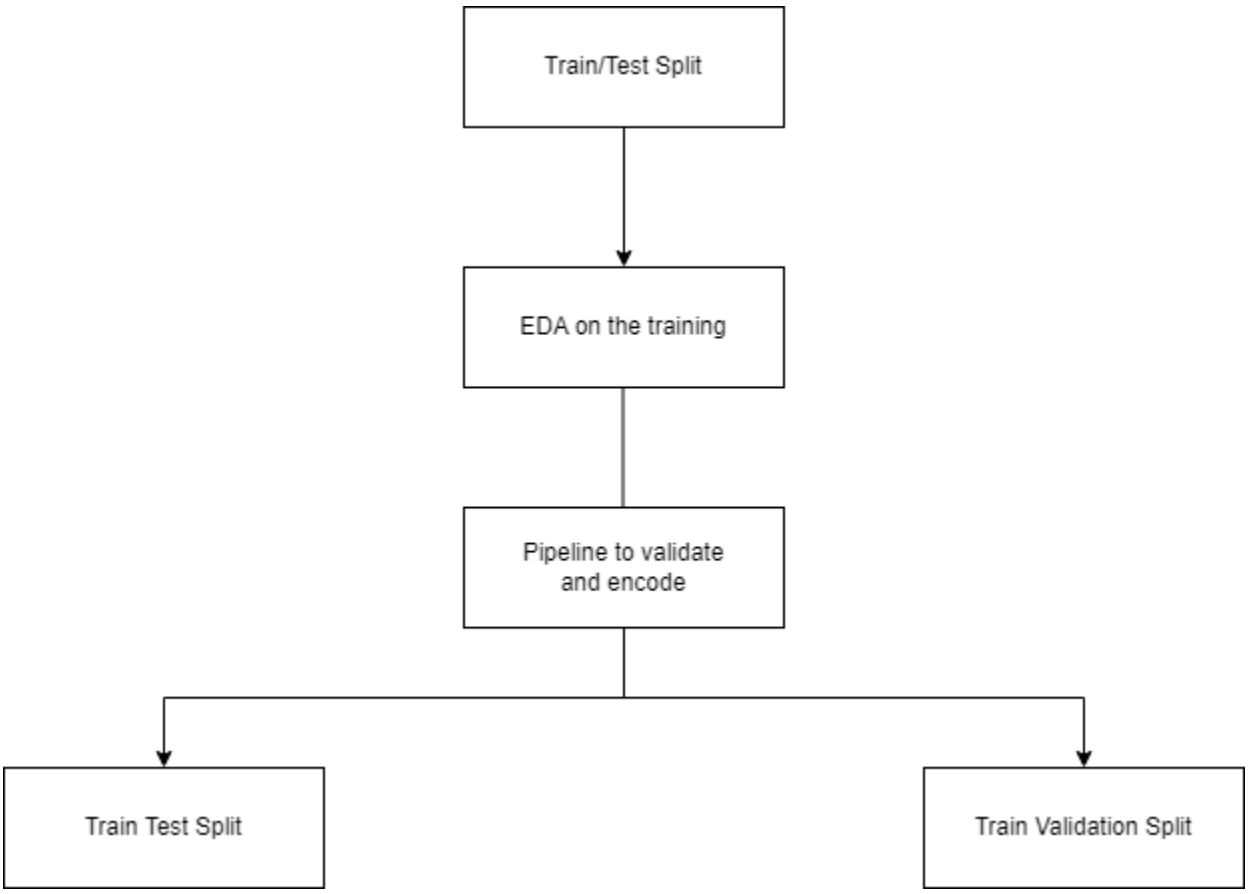
The same issue is present for every row in the data.

To prevent any analysis done on the test set and possibly any leakage that would result from that, the following steps of the cleaning and validation process were done after the train/test split, after the cleaning on the train a pipeline would be created to transform the test data as per the standards of the train set.

The process that was followed during the Cleaning and preliminary EDA for the purposes of modelling was to analyze every column for the existence of any abnormalities or unclean data while also gaining an understanding of the characteristics and distribution of the data.

After the EDA and Cleaning of the train set as mentioned above, the pipeline was created to apply the same transformations to the test set, for the training however to gain a yet better evaluation of the model, the train was split again to a train/validation set and transformations were applied separately where they would potentially cause data leakage.

The process was done as followed:



After the splitting and applying the pipelines, The necessary transformations were applied as follows

-Train/Test Set

The transformers were fitted on and transformed the training set and then fitted on the test set.

-Train/Validation Set

The transformers were fitted on and transformed the training set and then fitted on the validation set.

Next was the model Selection and Evaluation metric selection. starting with the evaluation metric, the recall and ROC_AUC score were chosen for the following reasons:

-For Recall

Recall was selected because of the problem we are dealing with here as the cost of False Negative errors (Customers churning when they were predicted not to) Is higher than the cost of False Positives (Customers not churning when they were expected to.

-For ROC_AUC score

The Receiver Operating Characteristic – Area Under Curve Score was used to measure the overall performance of our binary classification model as it is used to measure a model's ability to distinguish between the positive and negative classes.

It was also used because its more robust to class imbalance which the data suffers from.

And Lastly for the model selection, Logistic Regression, XGBClassifier and LightGBM were chosen.

The idea behind these selections was one linear model, one non-linear / non-parametric model.

However, LightGBM was also included because it's a faster and lighter version of XGBOOST.

Logistic Regression Performed Poorly on validation data so we didn't attempt to test it on the test data.

XGBOOST performed the best of all the models with recall of 0.897 and ROC_AUC of 0.94.

LightGBM performed the second to best with a recall of 0.84 and a ROC_AUC of 0.91.

Step-2: EDA

For the purposes of answering business questions and possibly uncover other insights or patterns in the data EDA was performed on the cleaned data extracted from the modelling process, preliminary investigations such as looking at the distribution of features was skipped as it was already performed during the modelling and as before the code is written with a thorough detail in the code notebooks attached with the documentation.

The standard steps that were followed during the EDA can be explained as followed.

Visual Investigation

Statistical Testing if applicable

Modelling for the purposes of uncovering relations

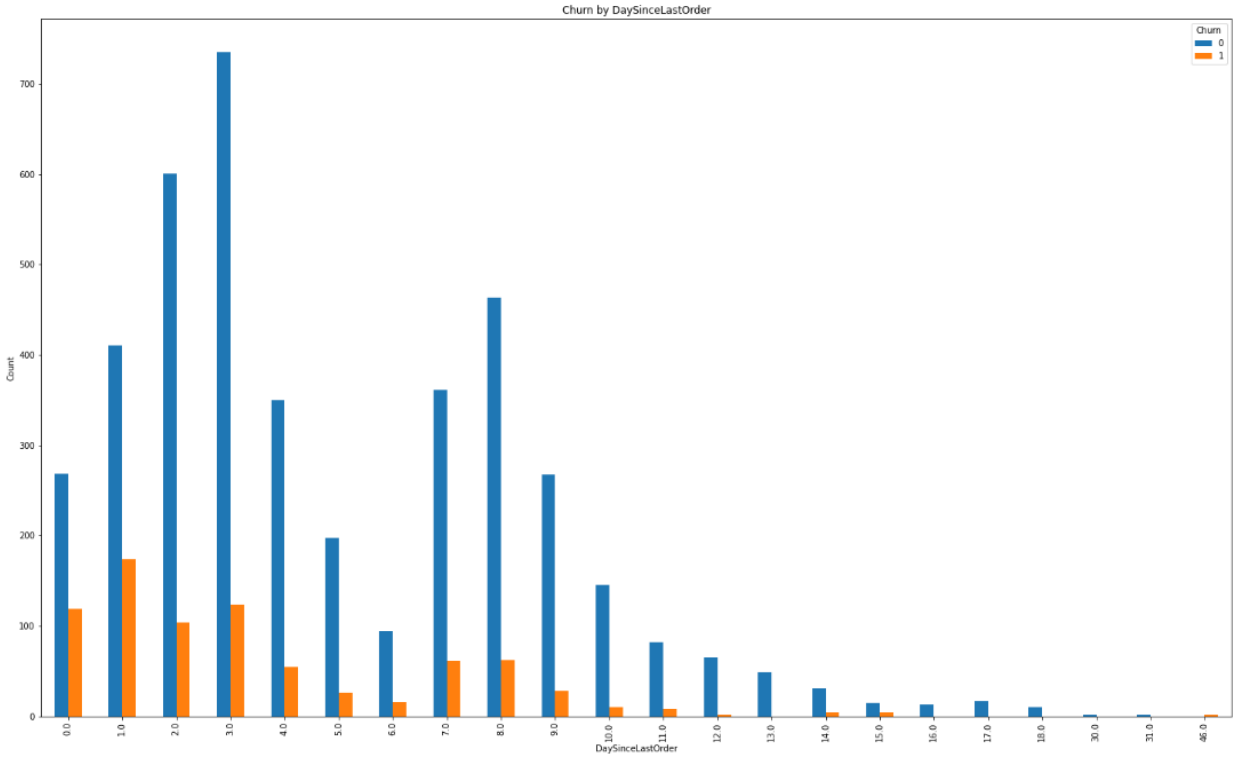
The steps above always followed the same order but were not always performed completely if sufficient evidence from the steps before it was achieved.

Will start first by tackling business questions and then move to the insights found.

Business Question 1:

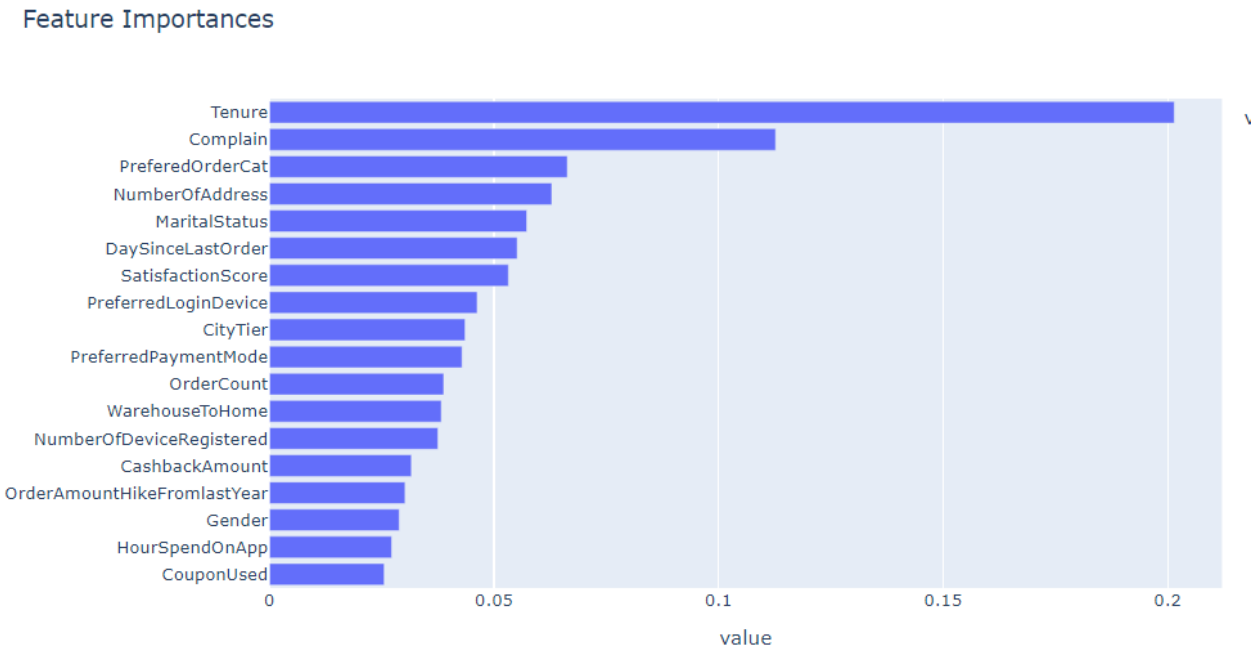
Analyze the number of days since the last order by the customer to create targeted marketing campaigns and offer personalized discounts.

When tackling this question other factors had to be considered first before segmenting customers by DaySinceLastOrder among which was, do the DaySinceLastOrder feature even relates to churn? so first the relation between them was first investigated to find a relation or the absence of one.



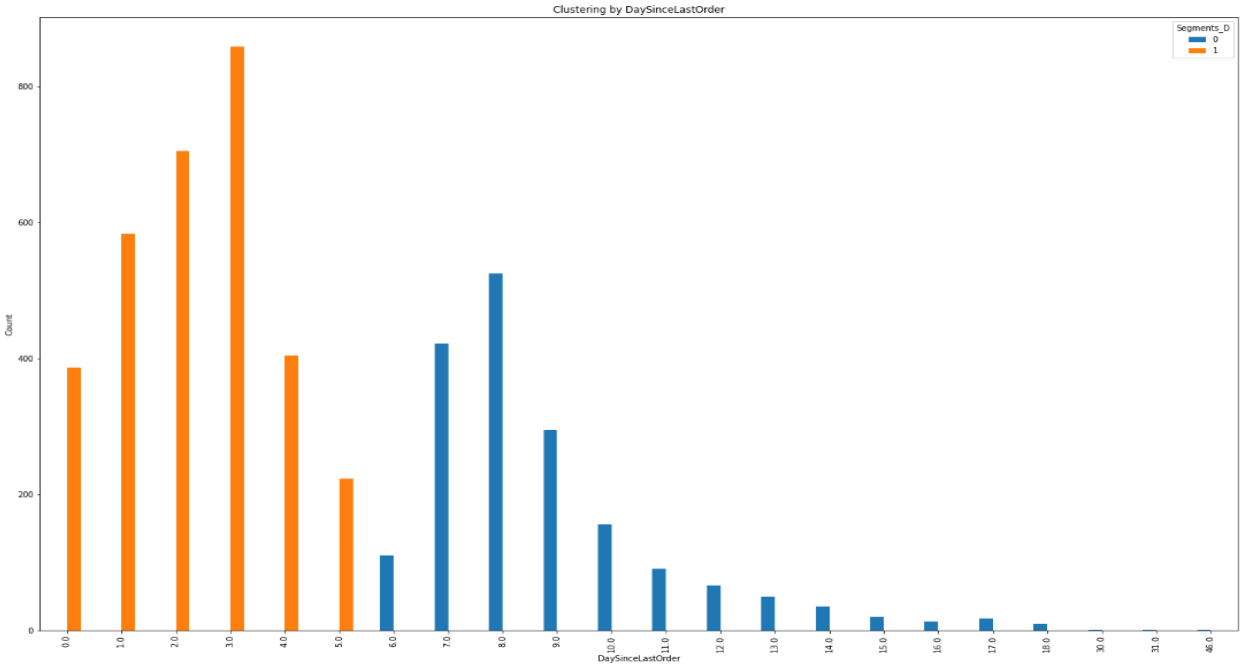
Looking at the distribution of the customers across the days by whether they churned, it seems that customers generally tend to churn less the more days pass since their last order.

However, despite the assumption the customers seem to churn as days goes by, the first 2 months seem to have an unusually large proportion of churning customers.

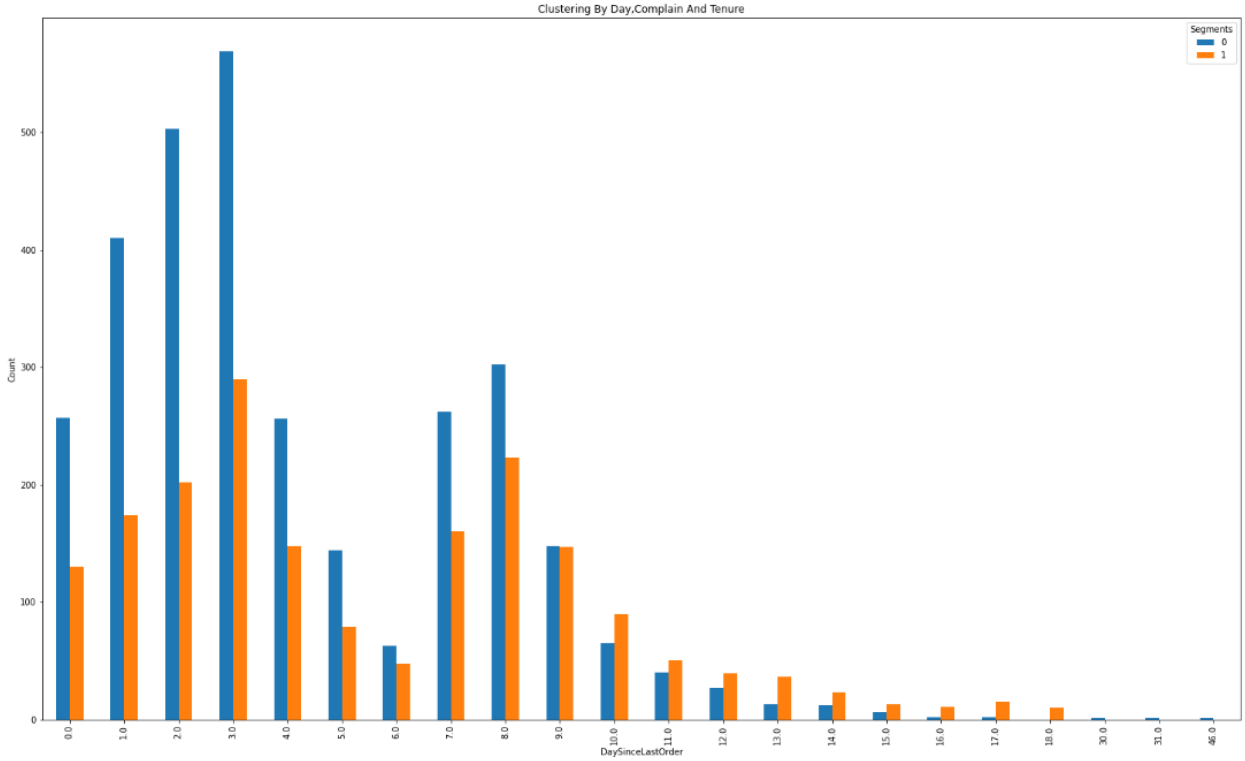


Further Supported by the feature importance extracted from the XGBOOST model, splitting with DaySinceLastOrder only would most likely provide weak results.

Other factors such as Tenure and complain need to be considered when considering Segmenting or clustering customers as they carry a huge effect as to whether the customer will churn or not, despite DaySinceLastOrder having an effect on churn, simply splitting the customers only by it would result in an inaccurate result.



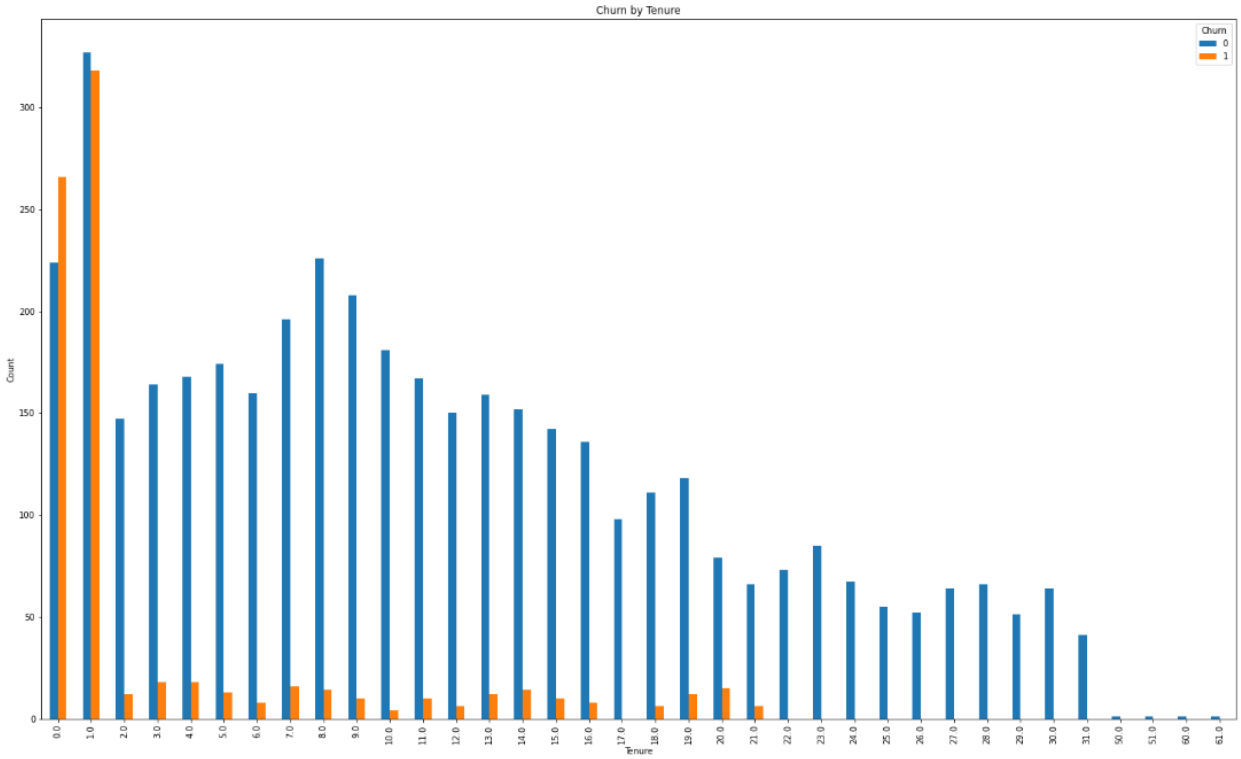
Clustering as it would be by DaySinceLastOrder Only



Clustering as it would be by DaySinceLastOrder, Complain and Tenure.

The reason for splitting the data into 2 clusters only was because of how volatile the churn rate in the early months with the first 2 months resulting in most of the

churn to the point that splitting customers by those who are In the first 2 months and those who come after would result in a very decent split.



Churn by Tenure

The splits in the first 2 graphs were done by using Kmeans clustering and while the Kmeans clustering performance using the 3 features reflects good results it doesn't capture the volatility found in the first 2 months of tenure which should be regarded with an utmost caution.

And with that in mind the following clustering of customers was considered in which Tenure has the highest weights, with respect to other variables as well.

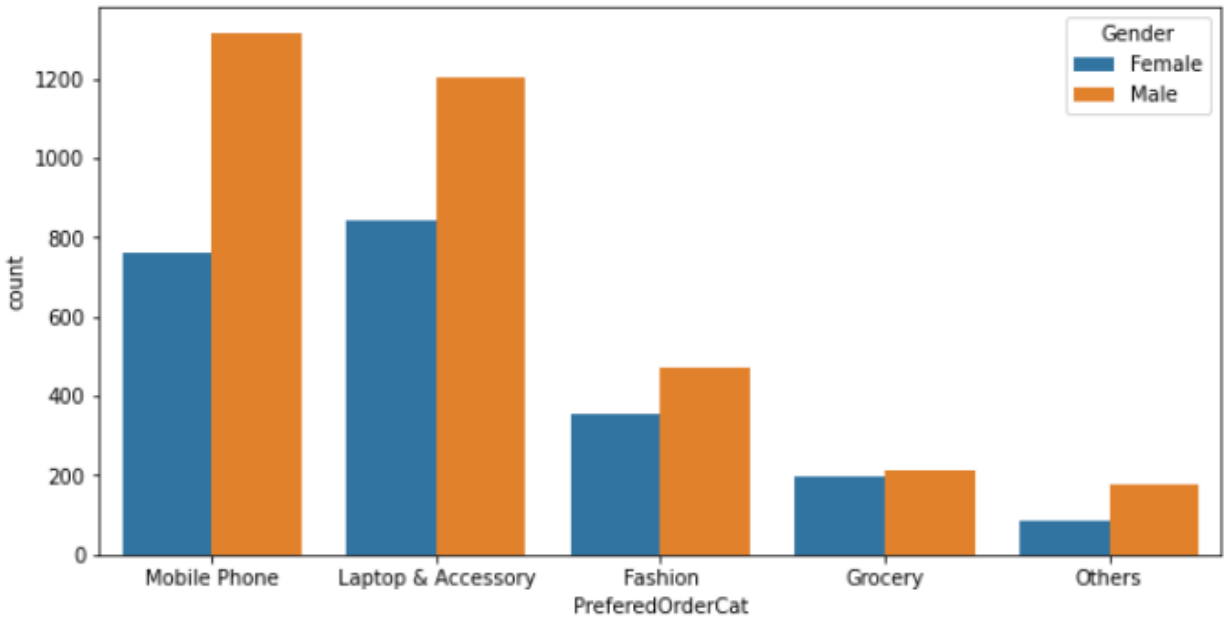
-customers in the first 2 months of their tenure would be clustered as volatile.

-customers in the months 3 to 22 would be according to whether they complained or not.

-customers in the months 22+ would be clustered as loyal customers and those have the least likely chances of churning regardless of whether they complained or not.

Business Question 2:

Is there any difference in the buying behavior of male and female customers?



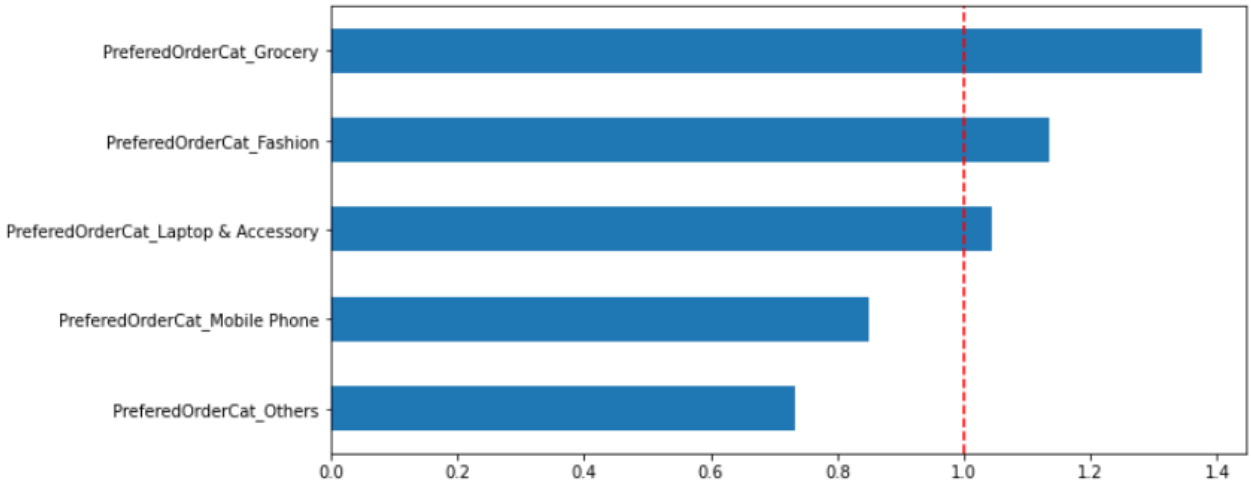
Preferred Categories by gender

Males are higher than females in all categories but that is due to males being higher in number in this dataset in general, what we are looking for is differences in proportions which makes the graph not the best way to look at it.

```
In [56]: #Chi-Squared Test

         cont_table = pd.crosstab(df['Gender'],df['PreferedOrderCat'])

In [58]: stat, p_value, dof, expected = chi2_contingency(cont_table)

         stat, p_value, dof, expected

Out[58]: (31.055258960935124,
          2.9829554534127733e-06,
          4,
          array([[ 329.51971581,  163.56305506,  817.81527531,  829.78330373,
                   105.31865009],
                 [ 496.48028419,  246.43694494, 1232.18472469, 1250.21669627,
                   158.68134991]]))
```

Chi-Squared test

With a chi squared test < 0.05, it seems that there is a difference between the buying behavior between male and female customers.



Logistic Regression – Odds Ratios

The odds ratios extracted from the logistic regression model shows us that there is a difference between the buying behavior of male and female customers which can be summarized as follows

-females are 1.4 times more likely to buy groceries.

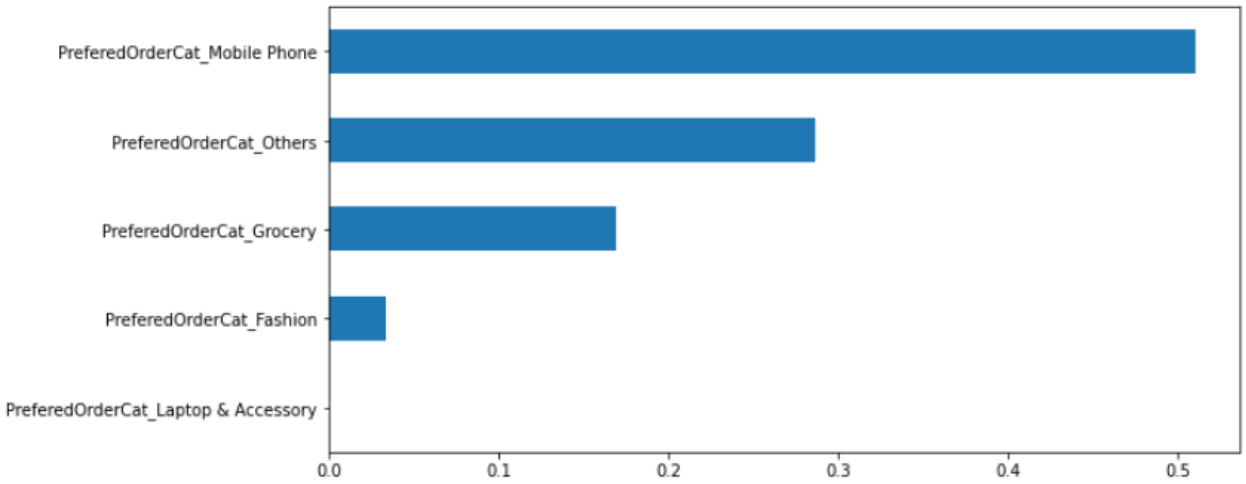
-females are 1.1 times more likely to buy fashion products.

-males are 1.4 times likely to buy products in 'Other's category.

-males are 1.15 times more likely to buy mobile phones.

-almost no difference in the buying behavior of males and females when it comes to laptop & accessories.

However, with a roc_auc score of 0.53 these results should be taken with a caution as a model this week is barely any better than random guessing, moreover it also speaks to the variable’s ability to split/predict the gender.



XGBClassifier – Feature Importance

While feature importance do not show us the direction of the effect, it shows us the strength of it which can be used to supplement the results of the logistic regression odds ratios and from these 2 charts we can conclude that:

Males are more likely to buy products of the Mobile Phone category.

Males are more likely to buy products of the Others category.

Females are more likely to buy products of the Grocery and fashion categories.

No difference between the males and females in the laptop & accessory category.

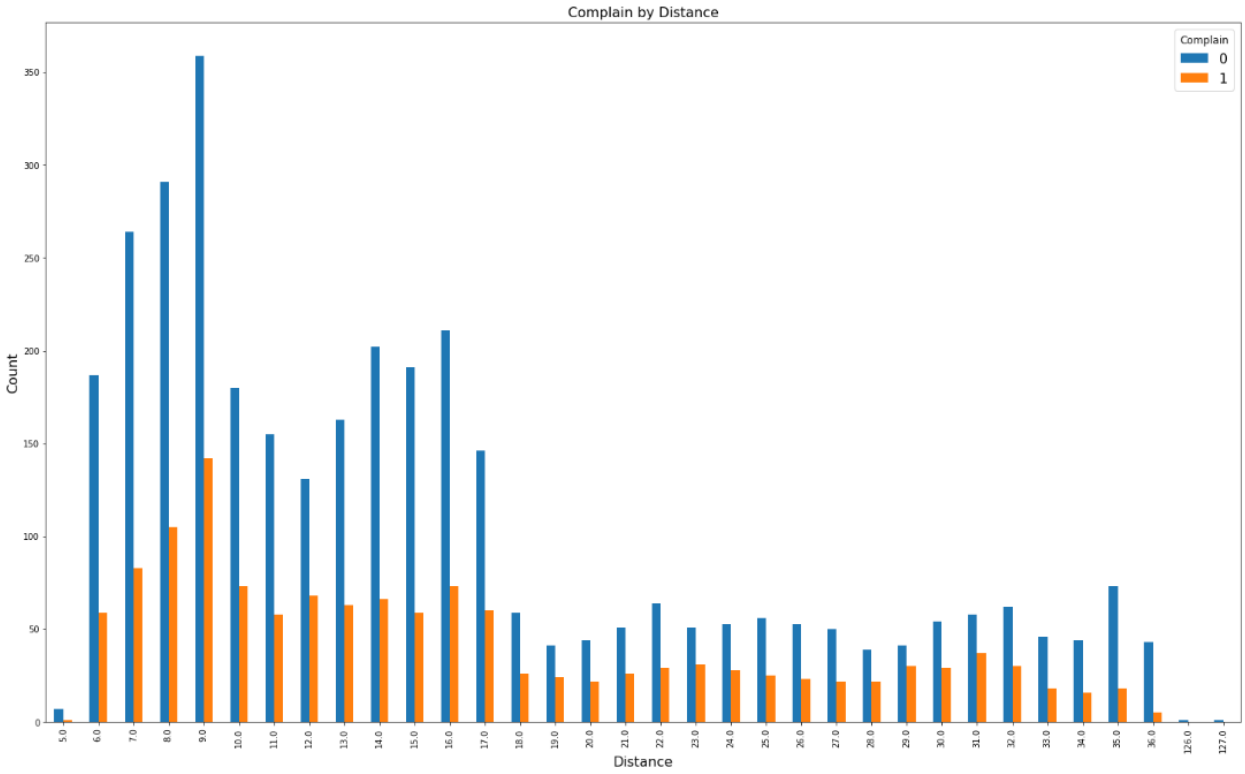
Business Question 3:

Provide key insights on why our customers churn and possible churn indicators.

For this question there is no specific answer as the specific reasons or indicators are present throughout the rest of the analysis.

Business Question 4:

Analyze the distance between the warehouse and the customer's home and check if it’s related to complains?



Complains by distance.

There is no apparent relation found in the graph as the distribution of both customers who complained or not seems to be the same across different distances.

```
In [80]: #ANOVA
         f_stat,p_value = f_oneway(df[df['Complain'] == 0]['WarehouseToHome'].dropna(),df[df['Complain'] == 1]['WarehouseToHome'].dropna()
         f_stat,p_value
Out[80]: (5.262816698769086, 0.021825542931292602)

In [81]: #Knedall's Tau
         df.corr(method='kendall').loc['Complain']['WarehouseToHome']
Out[81]: 0.03518265477429273
```

Statistical Testing (ANOVA & Kendall’s Tau,) The statistical test of ANOVA shows that there is possibly a relation however, both the f_stat and Kendall’s tau show there it is quite possibly a very weak relation.

Modelling (Logistic Regression & XGBOOST)

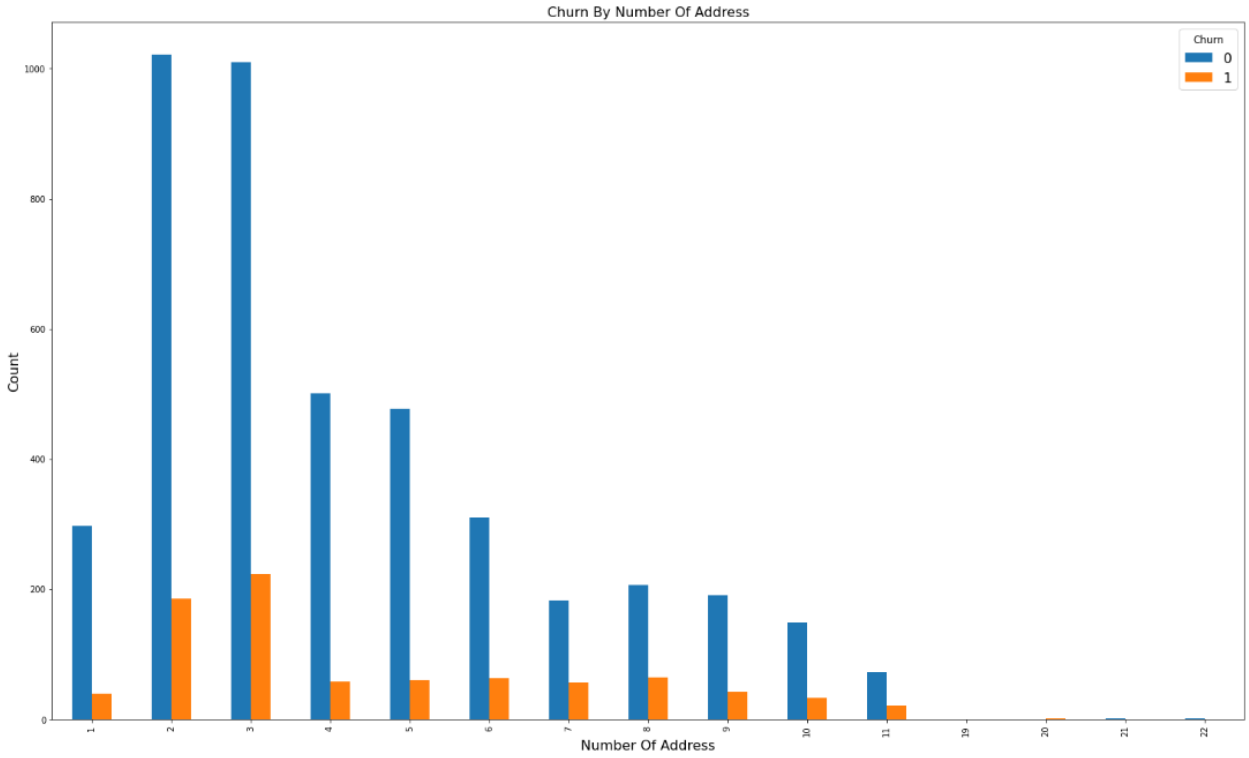
Logistic regression produced a model with a roc_auc score of 0.52.

XGBOOST produced a model with a roc_auc of 0.54.

Again, with such roc_auc score these models need to be interpreted with caution and whether the effect is significant or not will depend on the exact context of the business which we do not have a complete knowledge of.

Business Question 5:

Does the number of addresses added by customers impact the churn rate?



Churn by number of addresses

The visual representation of the relation shows no obvious relation between the number of addresses and the likelihood of churning.

```
#Chi Squared
cont_table = pd.crosstab(df['Churn'],df['NumberOfAddress'])
```

```
stat, p_value, dof, expected = chi2_contingency(cont_table)

stat, p_value, dof, expected
```

```
(67.88287739085857,
 4.654564206372349e-09,
 14,
 array([[3.08529663e+02, 1.13848277e+03, 1.06280568e+03, 4.88990409e+02,
         4.74852931e+02, 3.17677442e+02, 2.12893783e+02, 2.32852575e+02,
         1.98756306e+02, 1.61333570e+02, 8.14984014e+01, 8.31616341e-01,
         8.31616341e-01, 8.31616341e-01, 8.31616341e-01],
        [6.24703375e+01, 2.30517229e+02, 2.15194316e+02, 9.90095915e+01,
         9.61470693e+01, 6.43225577e+01, 4.31062167e+01, 4.71474245e+01,
         4.02436945e+01, 3.26664298e+01, 1.65015986e+01, 1.68383659e-01,
         1.68383659e-01, 1.68383659e-01, 1.68383659e-01]]))
```

Chi-squared test

With a p-value < 0.05, we believe that there is a relation between the 2 variables, but we can't speak of its practical significance.

```
#mutual information test

X= df['NumberOfAddress'].to_numpy().reshape(-1,1)
y=df['Churn']

mi_score = mutual_info_classif(X,y,random_state=42)
mi_score
```

```
array([0.00891859])
```

Mutual information test

With a mutual information of 0.008 we believe that the relation is vey weak between the variables.

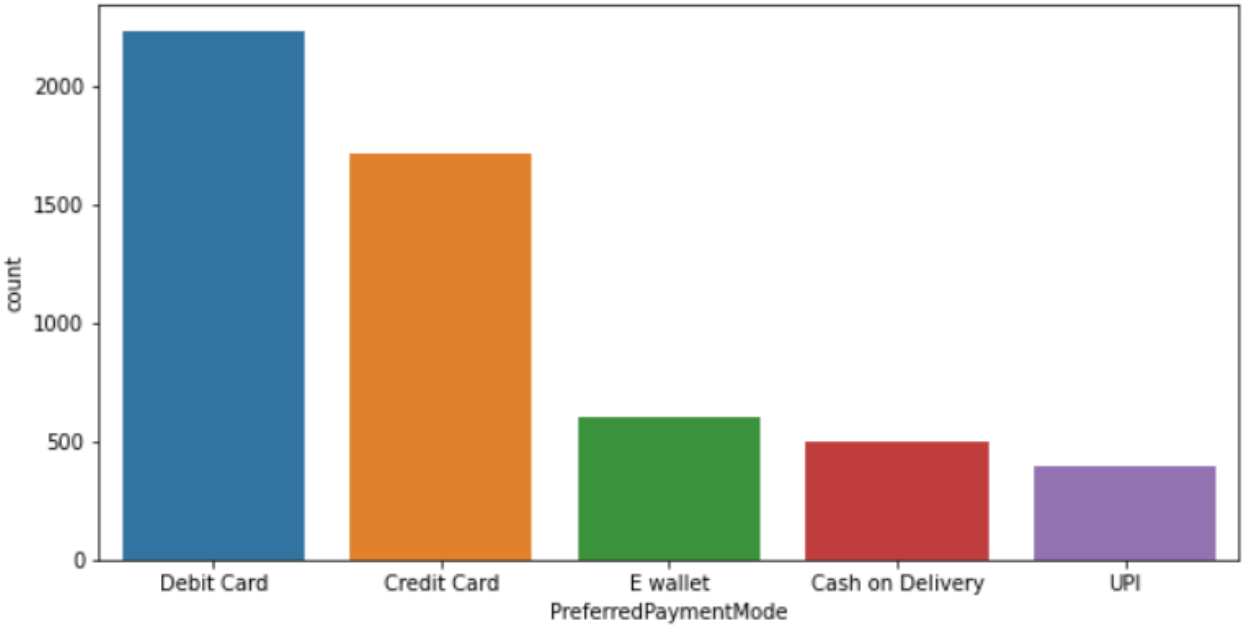
Modelling (Logistic Regression & XGBOOST)

Both models produced a roc_auc score of 0.55 which is slightly better than random guessing.

It seems that the effect of number of addresses on the likelihood of churning is very weak almost and has no practical significance.

Additional Insights

Preferred Payment



Count of Payment Methods

One thing of note in this graph is that the payment method UPI is that its newly rising payment method that is easy to integrate and takes few to no taxes which makes it idea to small and medium businesses, unlike banks which usually charge 2% per transaction, encouraging people to switch to such payment method with promotions and or sales could be potentially beneficial in the long run.

-Benefits of UPI

UPI payments are very fast and typically payment can be completed within seconds.

Almost every bank allows UPI transactions through mobile applications.

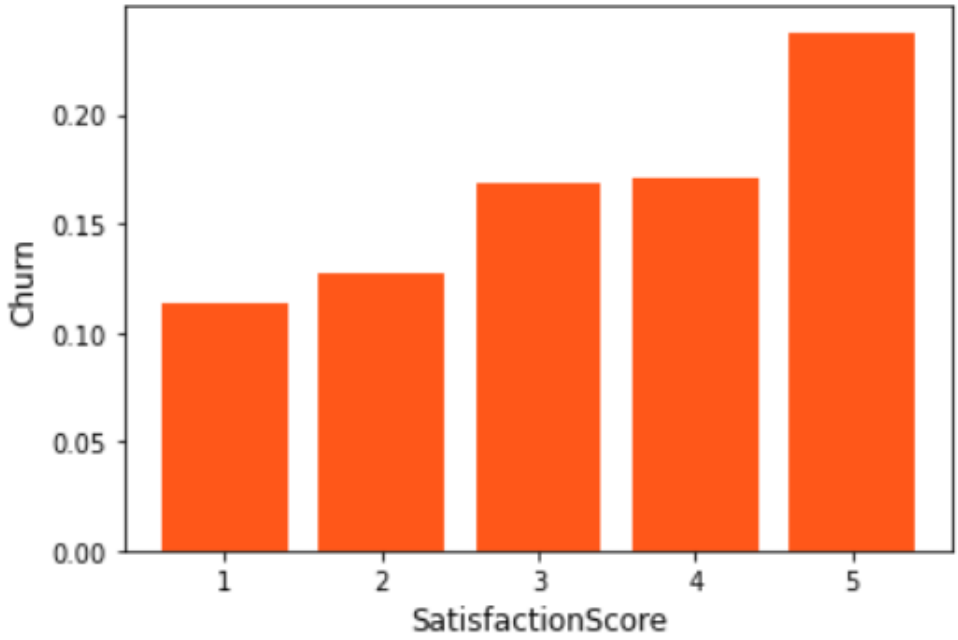
Payments are completely safe. To complete a payment, the user needs to have the SIM card of his mobile number present in his phone and needs to enter the secret MPIN each time.

UPI Payment facility allows individuals to request money from some other individual, which is not an option with other payment methods like IMPS, NEFT.

Payments can be made 24×7.

Free.

Satisfaction Score

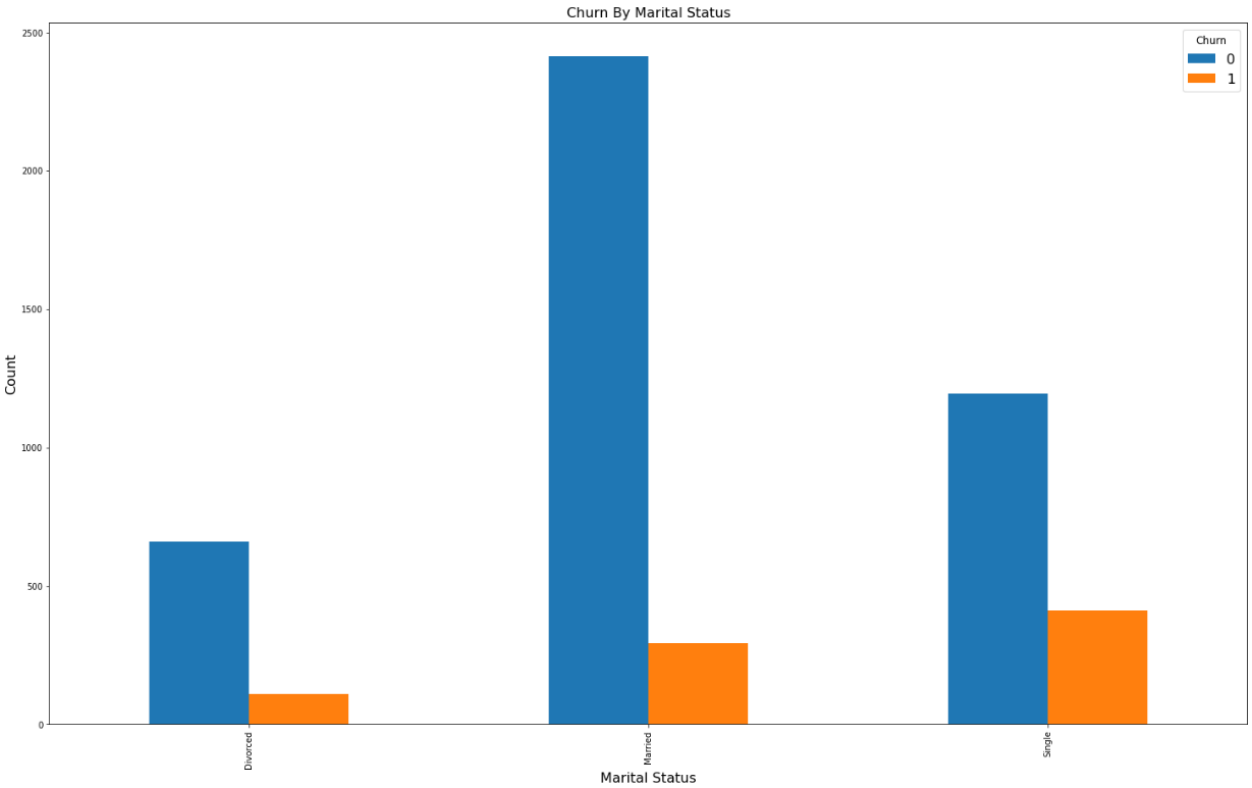


Somehow as the satisfaction score gets higher the proportion of customers who leave gets higher which can be possibly explained by some factors.

1- Satisfaction score is a weak variable and the effect we see here is a product of another variable in the data such as tenure which is further supported by the feature importance extracted from the xgboost model, although while This could potentially explain why satisfaction score is a weak variable, it doesn't explain exactly why we see such a high proportion of churning customers at satisfaction level 5.
2- The scales of this variable are reversed which would mean 5 translates to very unsatisfied, while 1 means very satisfied.

Regardless of the case without a deeper context of the data satisfaction score as it is can't be explained properly.

Marital Status



Churning by marital status

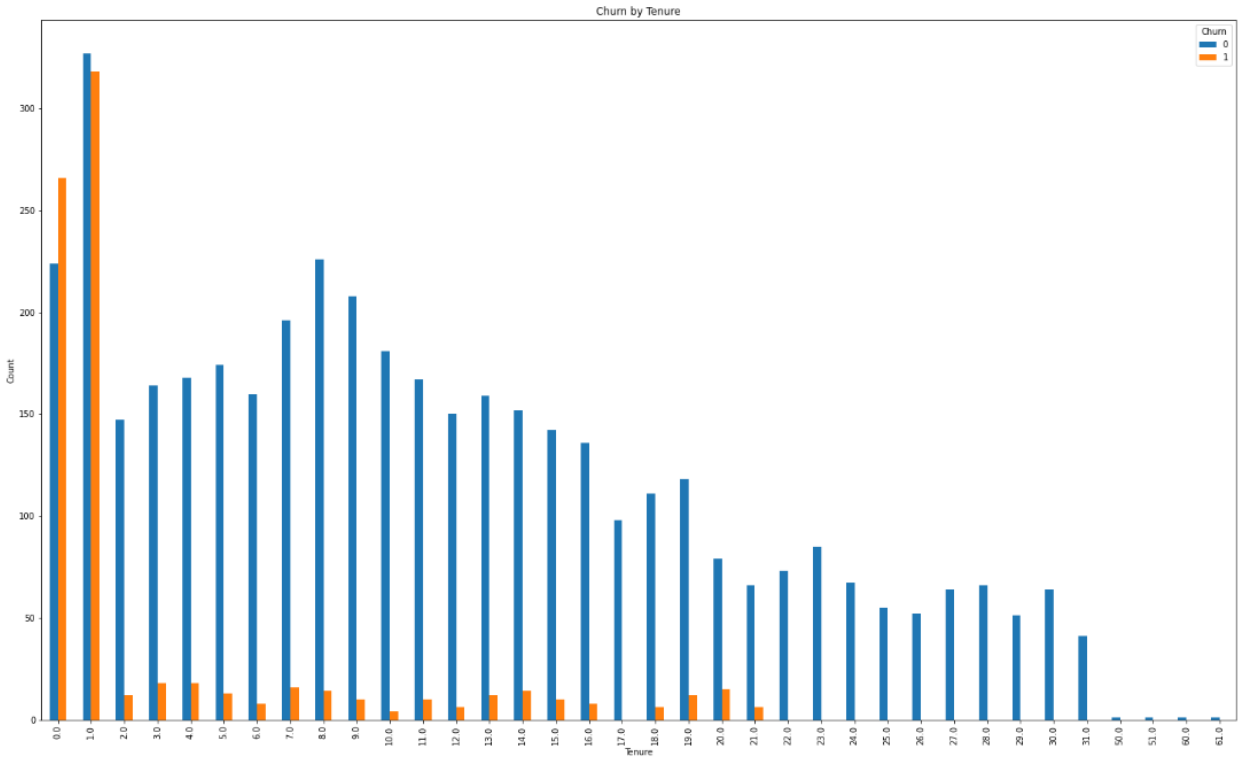
Single customers have an unusually high customer churning relative to how many single customers are there in the dataset.

Step-3: Presentation

The presentation will be attached with the deliverables.

Limitations:

Tenure

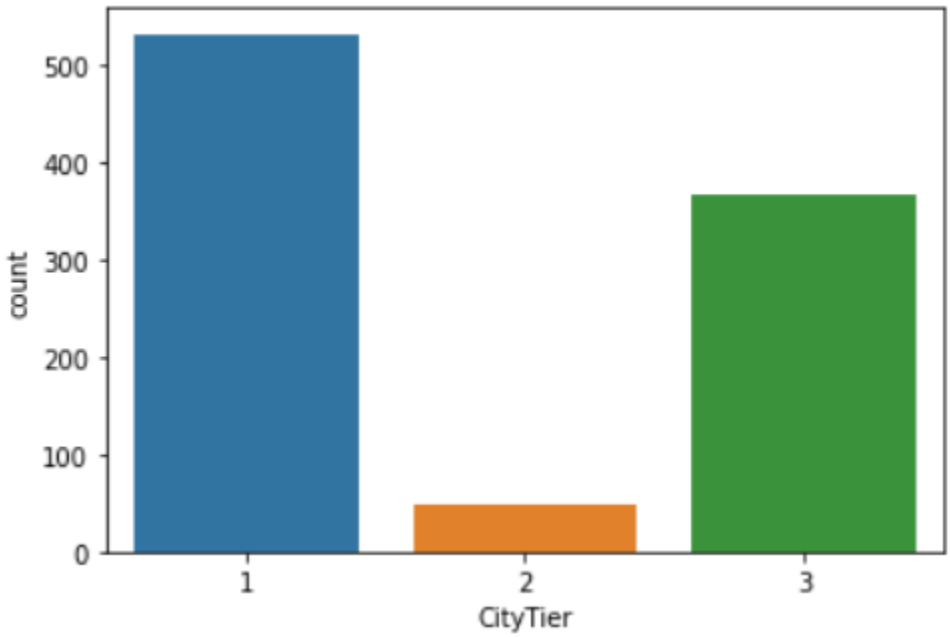


While tenure is our strongest variable in this dataset, there are really no reasons to explain why the customer churning is so skewed towards the right.

The closest explanation was due to the most preferred categories being phones and laptops but as these are not items that people usually buy on daily or semi daily basis.

Furthermore, when looking at the those whose preferred categories were phones and laptops no different patterns were found from those who had other categories as their preferred.

City Tier



Churning by CityTier

No sufficient explanation was provided to explain what exactly city tier is, internet searches each displayed different results depending on the locale so unless there is information about where the e-commerce store operates, no conclusions about the variable or its relation to churn can be drawn.