Predicting Customer Churn in Telecoms for proactive retention

By Elliot Ledson

What is logistic regression? Why use it?

- Logistic Regression is a statistical analysis method predicting a binary outcome, such as yes or no, based on prior observations of a data set.

- Logistic regression is commonly used for prediction and classification problems. Some of these use cases include: Fraud detection, Hotel Booking, Gaming, Text editing, and Credit scoring : Logistic regression models can help teams identify data anomalies, which are predictive of:

    - Detecting signs of fraud
    - Whether the user will change their journey.
    - What game a user may want to buy best on their behaviours/habits.
    - Toxic speech detection
    - Whether to extend a line of credit

What causes customer churn?

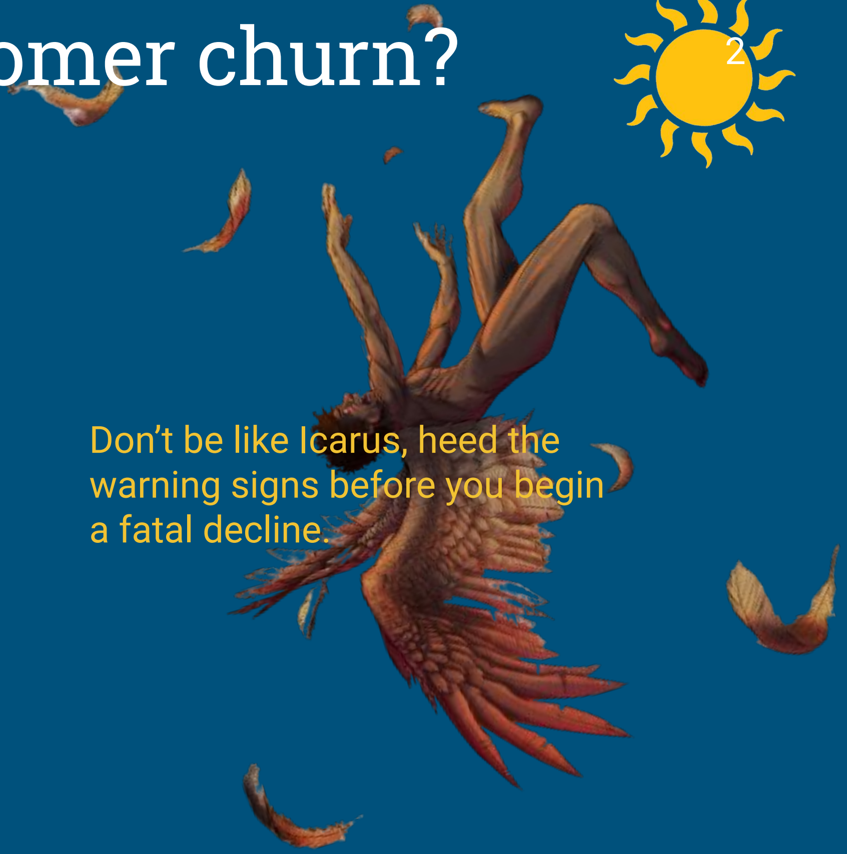
You're attracting the wrong customers

Your customer support needs work

Your product has bugs customers believe you can't fix

Your customers no longer see the value in your product

Your customers think your product's too expensive (or too cheap)



Don't be like Icarus, heed the warning signs before you begin a fatal decline.

What industry did I choose and why are they my focus?

Telecoms!

I am focussing on the Telecoms in industry.

Mobile/Cell phone communication is one of the worlds most successful industries and relies heavily on data to drive decisions.

Have you ever tried to cancel your Phone contract just to be passed to someone else to try and convince you to stay?

My Use Case

Build a churn prediction model using the Churn_Modelling.csv

By identifying which customers are most likely to churn, we can help our client take proactive measures to retain them and improve customer satisfaction.

My goal is to predict which customers are most likely to churn based on this information.

Domain Mastery

The telecom industry faces a high rate of customer churn, which is when customers switch to a competitor or cancel their service entirely.

Churn prediction models can help companies identify which customers are most likely to churn and take proactive measures to prevent it, such as offering special promotions or providing better customer service.

My Process

The steps I take

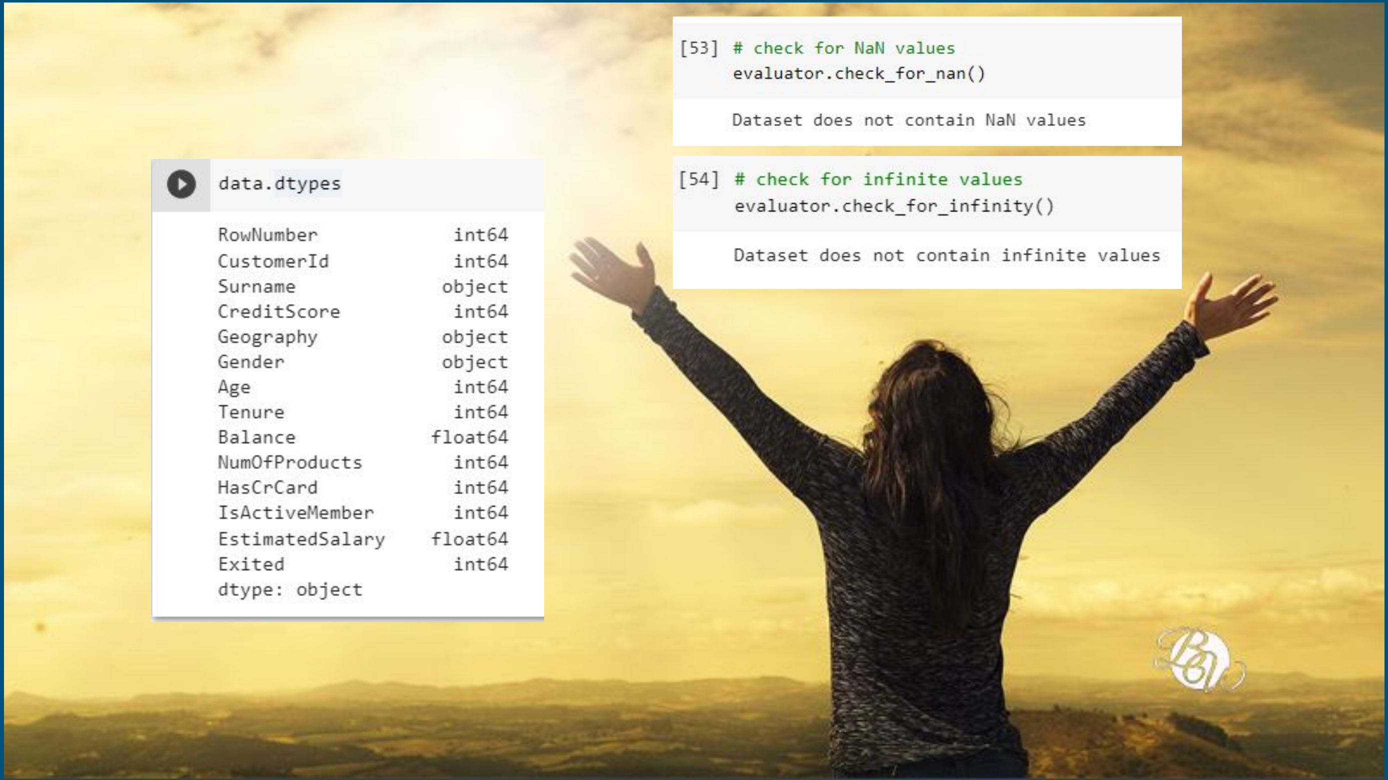
- Write a design document.
- Check the Data is pre-processed.
- Apply any preprocessing
- Train the model
- Evaluate the model
- Fine-tune the model

Why check the data is preprocessed?

To begin as you wish to proceed

We’re told the data is preprocessed, Learning through experience, I know you can’t take everything at face value, this 7 second video sums up my work ethic when a client hires me, telling me something has been done prior to my involvement.

The results are in



```
data.dtypes

RowNumber            int64
CustomerId           int64
Surname             object
CreditScore          int64
Geography           object
Gender              object
Age                  int64
Tenure               int64
Balance            float64
NumOfProducts        int64
HasCrCard            int64
IsActiveMember       int64
EstimatedSalary    float64
Exited               int64
dtype: object
```

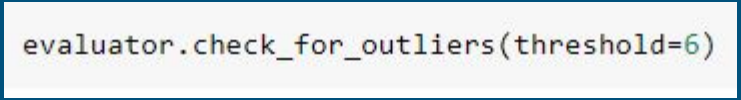
```
[53] # check for NaN values
     evaluator.check_for_nan()

     Dataset does not contain NaN values
```

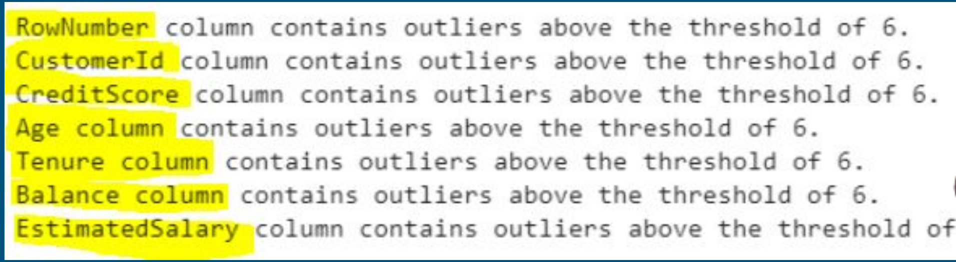
```
[54] # check for infinite values
     evaluator.check_for_infinity()

     Dataset does not contain infinite values
```

Evaluating Outliers...

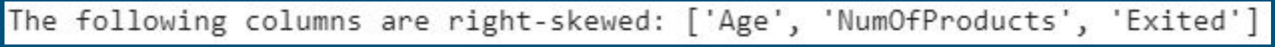


```
evaluator.check_for_outliers(threshold=6)
```



```
RowNumber column contains outliers above the threshold of 6.
CustomerId column contains outliers above the threshold of 6.
CreditScore column contains outliers above the threshold of 6.
Age column contains outliers above the threshold of 6.
Tenure column contains outliers above the threshold of 6.
Balance column contains outliers above the threshold of 6.
EstimatedSalary column contains outliers above the threshold of
```

Skewness



```
The following columns are right-skewed: ['Age', 'NumOfProducts', 'Exited']
```

Age

Age is logically right skewed to me as generally speaking the customers that stay will be older, or the accounts that stay will be older, so it is positive to have older accounts.

NumOfProducts

The num of products being right skewed indicates there is a significant amount of accounts with more products/packages/subscriptions.

Exited

I interpret Exited as The amount of accounts leaving. It is possible that this company is selective about which clients they choose to maintain.

This indicates they are great at drawing people possibly with new account promotional offers.

My hypothesis

I have gleaned from the data presented to you, that this company puts emphasis on acquisition but is very selective on the treatment of their acquisitions in regards to retention.

A high degree of early account closures mean there is an enticing reason to sign up. And a correlation with account age and the amount of products accrued It stands to reason this company retains more future business if they shift their emphasis to offering newer accounts more benefits earlier.

Tincy wincy bit of preprocessing

What do Gender and Geography have in common?

I ended up needing to label encode and one hot encode these as they are String Objects.

What results did I end up having?

I implemented a confusion matrix however I am having trouble where I encounter false results.

I was frequently identifying negatives but not finding any positives.

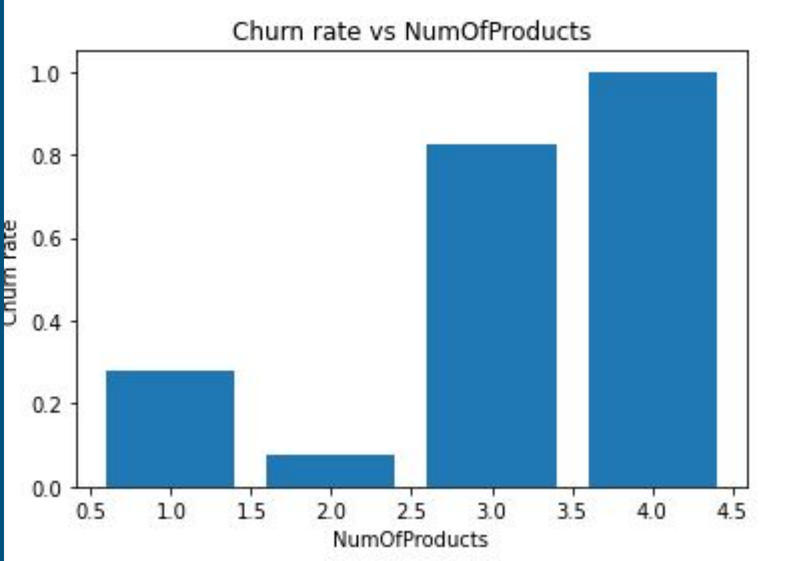
I found out in the 8000 Feature dataset there was over 2000 counts in exited.

Nothing to see here

So the main problem I am encountering is that the current functionality of my code, must be missing something or the data itself is unbalanced?

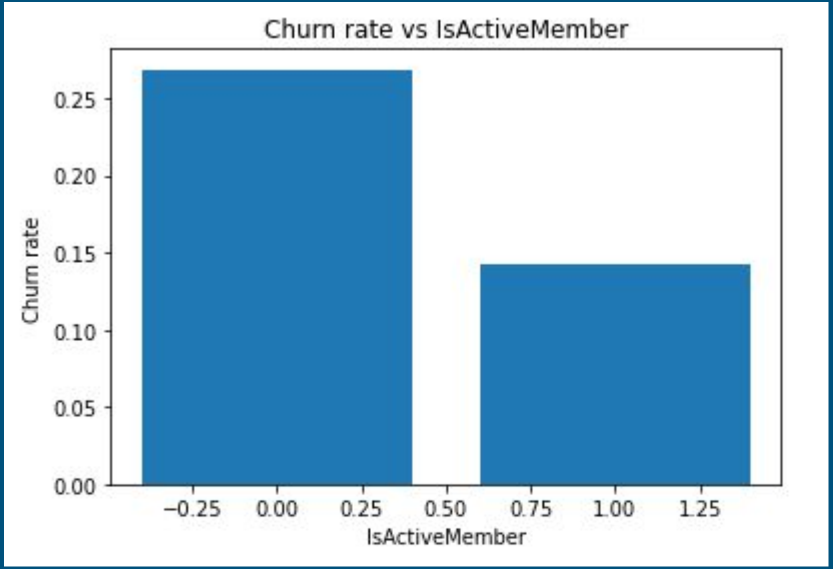
No matter how much I adjusted the threshold epoch and eta, my positives were low and my negatives are high. Despite an accuracy rating of over 75% consistently.

Products owned and Churn



Could it be that people have left and comeback because of increased benefits?

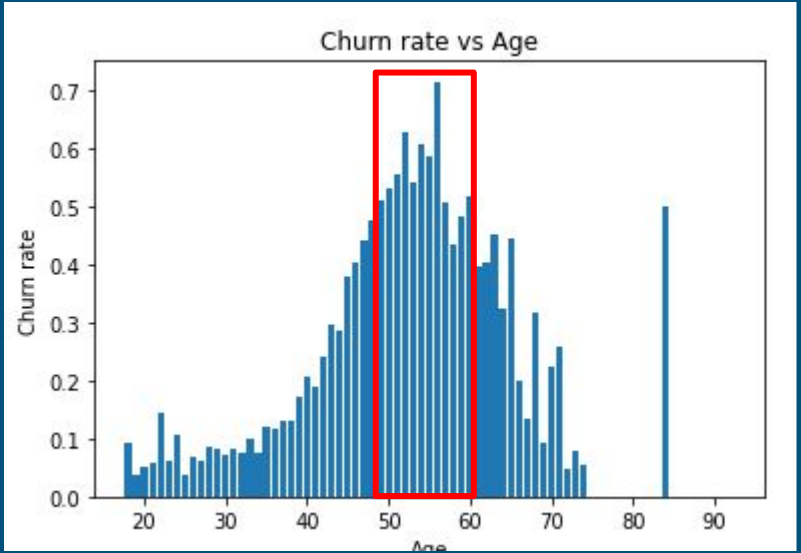
Less active members churn more.



There is a clear separation here for churn rate happening for active members in comparison to less active members. Which backs up my initial hypothesis.

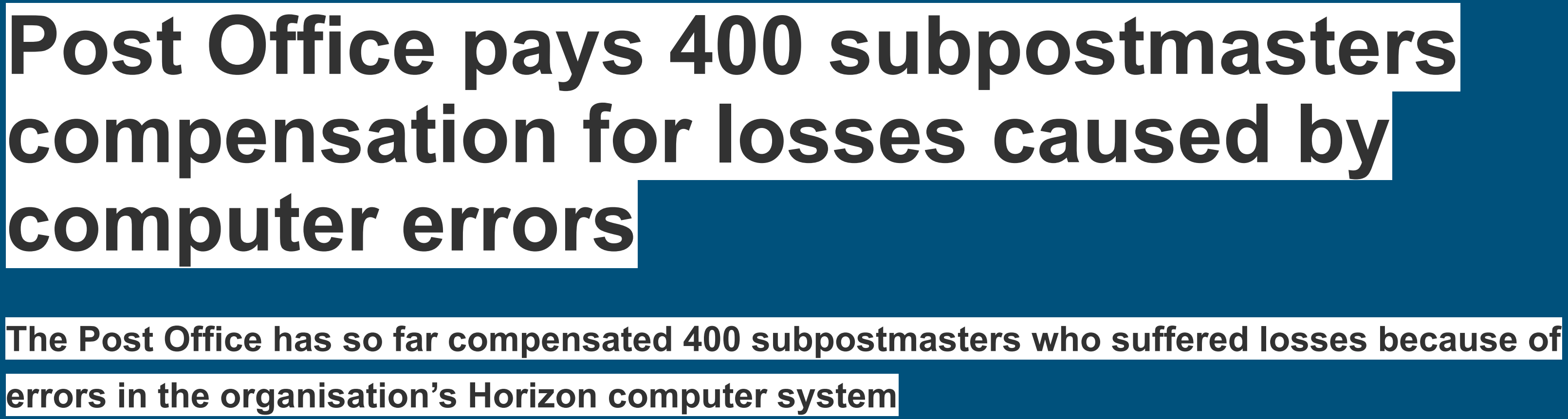
Your service isn’t engaging enough?

The age drop off



Your product seemingly has a higher churn rate for people between 50 and 60? It might be that the level of service you are providing isn’t appealing to that age demographic.

It is entirely possible the data is the problem?



Post Office pays 400 subpostmasters compensation for losses caused by computer errors

The Post Office has so far compensated 400 subpostmasters who suffered losses because of errors in the organisation’s Horizon computer system

Considerations to improve the model in the future

Investigate if the classes are unbalanced and if they are use oversampling techniques such as SMOTE (Synthetic Minority Over-sampling Technique) to balance the classes in the training data.

Consider feature engineering, such as creating new features from existing ones, to capture more information about the problem.

In conclusion...

My tests could be wrong however, they back up my initial hypothesis rather well. Other metrics to consider are nationality and gender too, my research showed a significantly higher amount of women contributing to Churn and a high concentration of German residents, so it could indicate German Women between 50 and 60 are looking for a service that appeals to them more. That could be something to feed back to the marketing department.

Otherwise it may be worth evaluating whether a loophole is being exploited for additional products/services.

It is definitely worth investigating technical error.