**Ideas:**

**Can you predict if a customer will make a purchase on a website using machine learning?**

Apply machine learning techniques and analysing data to investigate online shopping based on clickstream data.

Using machine learning models to predict if a customer will make an online purchase or not? What methods could encourage customers to come back and make the purchase?

Build dashboard in order for the company to see the data continuously when its updating.

Use data to analyse the data and gain insight for now and also use machine learning to predict the future visits.

Using clickstream data, more businesses have gone online in comparison to instore. Analyse their behaviour instore and try to prevent it.

Try all various models for classification as it will be 0 or 1/ true or false if they will make a purchase it or not?

Analyse the data where they do not make a purchase to find the common characteristics, and this could be a reason as to why they abandon their cart.

* Use different machine learning models, others haven’t used as many models.
* coming up with preventative models e.g free shipping pop up notification, free shipping in email, free returns, % discount,

**Motivation:**

* Been on websites, decided not to buy anything then get emails with offers or reminders to finish off the payment and then other times I get nothing
* Wondered what exactly trigger this, why do I get emails and not other times

**Dataset:**

Can only use google analytics etc for your own website – in order to have large number of rows using dataset

[**https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset**](https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset)

UCI Machine Learning Repository Dataset – uses google analytics

Over 12000 sessions in each one is a different user

**Columns:**

* Administrative: amount of pages viewed by visitor regarding account administration
* Administrative\_Duration: the total amount of time spent by a visitor (in seconds) on the pages regarding account administration
* Informational: amount of pages viewed by visitor regarding information ecommerce, corresponding and purchase procedure
* Informational\_Duration: the total amount of time spent by a visitor (in seconds) on pages regarding information e-commerce, corresponding and purchase procedure
* ProductRelated: amount of pages viewed by visitor regarding productrelated pages
* ProductRelated\_Duration: the total amount of time spent by a visitor (in seconds) on the pages regarding product-related pages
* BounceRates: the average value of bounce rate of pages viewed by the visitor (Google Analytics measurement)
* ExitRates: the average value of exit rate of pages viewed by the visitor (Google Analytics measurement)
* PageValues: the average value of contribution unique pages viewed by the visitor (measured by Google Analytics)
* SpecialDay: proximity of website visits with special day
* Month: the month when the session visited by the visitor
* OperatingSystems: operating system used by the visitor
* Browser: browser used by the visitor
* Region: geographic region data based on IP-based session visited by the visitor
* TrafficType: traffic source type which brings the visitor to the Web site (e.g., direct, search engine, display ads, SMS, etc.)
* VisitorType: visitor category of session visited by the visitor as ‘‘new visitor,’’ ‘‘return visitor,’’ or ‘‘other’’
* Weekend: time of session visited by visitor wheatear “weekday” or “weekend”
* Revenue: the class label indicating whether the visit was completed with a purchase or not

Rates explained:

<https://support.google.com/analytics/answer/2525491?hl=en>

"Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration" represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories. The values of these features are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, e.g. moving from one page to another.

The "Bounce Rate", "Exit Rate" and "Page Value" features represent the metrics measured by "Google Analytics" for each page in the e-commerce site.

The value of "Bounce Rate" feature for a web page refers to the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session.

The value of "Exit Rate" feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session.

The "Page Value" feature represents the average value for a web page that a user visited before completing an e-commerce transaction.

The "Special Day" feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction.

The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date.

For example, for Valentina’s day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8.

The dataset also includes operating system, browser, region, traffic type, visitor type as returning or new visitor, a Boolean value indicating whether the date of the visit is weekend, and month of the year.

* use machine learning model on the data to see which model will work best for predicting
* analyse data and make graphs and charts in beginning of data
* Calculate corelation coefficient
* Predicting revenue will be true / false using all the other columns

In a regular retail shop the behavior of customers may yield a lot to the shop assistant. However, when it comes to online shopping it is not possible to see and analyze customer behavior such as facial mimics, products they check or touch etc. In this case, clickstreams or the mouse movements of e-customers may provide some hints about their buying behavior. In this study, we have presented a model to analyze clickstreams of e-customers and extract information and make predictions about their shopping behavior on a digital market place.

Research Question?

This research evaluated the accuracy of predictive models in determining if a customer will abandon their cart while online or finish their purchase. Multiple methods were utilised to analyse the chosen data in order to determine and enchance predictive accuracies. The overarching research question is as follows:

By using clickstream data can one predict if someone will make a purchase on an e-commerce website, or will they abandon their purchase?

Research objectives:

Five research objectives were identified, namely:

* To identify Machine Learning algorithms which will best predict online expenditure.
* To investigate the correlation between clickstream data features and online purchasing.
* To utilise the clickstream data features to increase the accuracy of the prediction model.
* To evaluate the model performances.
* To research methods in preventing cart abonnement – which pages did they visit vs which did they not?

Literature review: 4/ 5 papers for each??

1. The move to online retail
2. Need for/ uses of clickstream data, expenses etc
3. Predictive Analysis Methods for determining cart abandonment/ expenditure online
4. Methods of enticing customers back to finish purchase?

Paper links:

Website about clickstream data: <https://www.techtarget.com/searchcustomerexperience/definition/clickstream-analysis-clickstream-analytics>

**Capturing evolving visit behavior in clickstream data**

<https://onlinelibrary.wiley.com/doi/pdf/10.1002/dir.10074?casa_token=wYhdAnN8NGMAAAAA:RjNBYk0IEI5Cl_L-vDoH97EorW9viVS3oGYJ1gIQwz7_p2eUz9jBK9aZvSpzgCSmb3vNOfUGtu8E9dc>

**Modeling Online Browsing and Path Analysis Using Clickstream Data** <https://pubsonline.informs.org/doi/pdf/10.1287/mksc.1040.0073?casa_token=6YDrwfJgP1QAAAAA:brdlAZbeV_uQ8jSJNDKqXbwu1oAurKsz_NQe8jkwKFhrcHZEpa5vVggR0z-HCIrKo20zLqQcHg>

**Visualization and Analysis of Clickstream Data of Online Stores for Understanding Web Merchandising**

<https://link.springer.com/content/pdf/10.1023/A:1009843912662.pdf>

**Modeling Consumer Purchasing Behavior in Social Shopping Communities with Clickstream Data**

<https://www.tandfonline.com/doi/pdf/10.2753/JEC1086-4415160202?casa_token=utdXAfCqWH8AAAAA:apRKc1UNc14VihM32cMOLY5ZCOL4Oqjg1bU-SZ8vYKVGNcMr_pPipo9UC6VSUutvRfZn3fueJbg>

**Predicting online shopping behaviour from clickstream data using deep learning**

<https://www.sciencedirect.com/science/article/abs/pii/S0957417420301676?casa_token=95QqUQ0rhXgAAAAA:NrHBgHwmAgLZcgw-lAQLusqlTGlIZC3RbvucXtDQuBmYwXHzNWMdHQlFM3bsdfQUkkrqvH19>

**Shopping Hard or Hardly Shopping: Revealing Consumer Segments Using Clickstream Data**

<https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9422915&casa_token=Erv0EeQ-LwMAAAAA:besEztd7EalcjfMOn_C8vtJqYO_C_gI7Uc-WaebF5zREcDbDSTFLHwdOxTrSFeTW-Q79Uki6>

* Good one

**Using Clickstream Data to Improve Flash Sales Effectiveness**

<https://onlinelibrary.wiley.com/doi/pdf/10.1111/poms.13238?casa_token=2UtaHrL9YC8AAAAA:1rRijFbtX60V7xQpTQr2ua16x7b2K5OPRSWsreTVzOHVkm7mLoyB4uNNm6JtCAPixlISNwhE9mU>

**Analysis and prediction of Ε-customers' behavior by mining clickstream data**

<https://ieeexplore.ieee.org/abstract/document/7363908?casa_token=_58xHayVW_sAAAAA:teZb008mvu_oqK5akS0PShEO3_twjA2K5MZ_kJqthwfFMW1q3d82Dk5Fb3KPPf1vPz-XhWdg>

**Predicting online shopping cart abandonment with machine learning approaches**

<https://journals.sagepub.com/doi/pdf/10.1177/1470785320972526?casa_token=_l7nDQnADp8AAAAA:5L_phVDE824RmsIaY-sp7CinmtPGna-RFEtWzWGpVwAY6yN8Xc9v0tYMfXCPBTj97FtfacG9EBk>

**APPLYING DATA MINING TECHNIQUES TO INVESTIGATE ONLINE SHOPPER PURCHASE INTENTION BASED ON CLICKSTREAM DATA**

<https://www.fortunepublishing.org/index.php/rbaf/article/view/15>

<https://www.stacktome.com/blog/a-guide-to-data-warehousing-clickstream-data> ->>>

**Going beyond charts and dashboards**

Tracking KPIs with charts and dashboards is helpful for monitoring business health and detecting problems in real-time. Though this is useful when making high-level business decisions. To truly bring business to the next level the data can be utilized for optimizing activity down to each customer level. One of the most popular examples is personalizing customer experience.

Research Motivation:

* Been on websites, decided not to buy anything then get emails with offers or reminders to finish off the payment and then other times I get nothing
* Wondered what exactly trigger this, why do I get emails and not other times
* Having used machine learning on other pojects, wondered how it relate to online shopping something that I do at least twice a month

This study purports to analyse the online shop visitors' behaviour patterns as powerful reference strategies to improve services, content, page display for better web modification, visitor behaviour predictions, or more marketing strategies. Thus, online shops can gather useful information regarding the number of loyal customers (Lee and Kwon, 2008), the more or frequently web page accessed by visitors, visitor behaviour patterns, pages that are rarely visited, future targeted customers for promotion and others

Research Objectoves

Five research objectives were identified, namely:

* To identify Machine Learning algorithms which will best predict online expenditure.
* To investigate the correlation between clickstream data features and online purchasing.
* To utilise the clickstream data features to increase the accuracy of the prediction model.
* To evaluate the model performances.
* To research methods in preventing cart abonnement – which pages did they visit vs which did they not?

Data Retrieval

Classification

Models

Machine Learning algorithms

* Classification
* Decision Tree
* Random Forest
* K-NN Classifier
* Logistic Regression
* XGBoost
* Adboost
* Neural Net
* Deep Learning
* Naïve Bayes
* Rule Induction

Google Colab:

Dataset is imbalanced

Separate files for each technique like smote, random under sampling etcc

Graphs- https://colab.research.google.com/drive/12XxbSUlD0g87awAFLAhoKBXLl7hlD9YI

Normal - https://colab.research.google.com/drive/1jsND4JZpkX9d4azQk6JzID4slo2pukfl#scrollTo=ET6e2YcuGF3m

Random under sampling - https://colab.research.google.com/drive/1TcPtajvnttoY42dUaKbxXWz-Wsf0SD\_9

Over sampling smote - https://colab.research.google.com/drive/1v6WcshxWugZZqGZ1JOOYqRHoqMNMoiuS

Under sampling near miss -

https://colab.research.google.com/drive/1N9vfiRqRgf5PSjwq5PjWaxXJt\_lgPGjy

Tomeklinks - https://colab.research.google.com/drive/1wGOgsaUjRLeZfwpc5yG6XPWXRJS5uvkf

Borderline smote - https://colab.research.google.com/drive/1Dcd8tmTpuj2jrgoaDHpGxA28MqNxqwLe

Adasy - https://colab.research.google.com/drive/1Ty5g-gJHx0-mUUmkS7QXkZrQSLA2UwU-

KNN:

knn\_params = { 'n\_neighbors':[2,3,5,7,9,11,13,15,17,19] }

Logistic Regression:

lr\_params = {'penalty':('l1', 'l2'),

'C':(0.01, 0.05, 0.1, 0.5, 1, 5, 10),

'solver': ['newton-cg', 'liblinear', 'lbfgs']}

Random Forest:

grid = {'n\_estimators': [10,50,100,200],

'max\_features': ['sqrt', 'log2'],

'max\_depth' : [4,5,6,7,8],

'criterion' :['gini', 'entropy'],

'min\_samples\_split': [2,5,10],

'min\_samples\_leaf': [1,5,12]}

Decision Tree Classifier:

params\_dt = {

'max\_depth':[1,5,10,12],

'min\_samples\_split':[2,4,6,8,10],

'min\_samples\_leaf':list(range(1, 16)),

'criterion': ["gini", "entropy"]

}

XGBoost:

params\_xgb = {

'max\_features': ['sqrt', 'log2'],

'subsample': [0.4, 0.6, 0.8],

'max\_depth': [1,5,10,15],

'n\_estimators':[10,40,60],

'learning\_rate':[0.1, 0.4, 0.8, 1.6],

}

AdaBoost:

abm\_param\_grid = {'n\_estimators': [10,50, 120, 200],

'learning\_rate':[0.01,0.1,0.5,1]}

Support Vector:

svm\_param\_grid = {'C': [0.1, 1, 10, 100],

'gamma': [0.01, 0.1, 1, 10, 100],

'kernel': ['linear', 'poly', 'rbf']}

Classification Report:

precision recall f1-score support

0 0.87 0.98 0.92 3127

1 0.67 0.20 0.31 572

accuracy 0.86 3699

macro avg 0.77 0.59 0.62 3699

weighted avg 0.84 0.86 0.83 3699

Take weighted average

Precision attempts to answer the following question:

- What proportion of positive identifications was actually correct?

Precision = TP/(TP+FP)

Recall attempts to answer the following question:

- What proportion of actual positives was identified correctly?

Recall = TP/(TP+FN)

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same.

Accuracy = TP+TN/TP+FP+FN+TN

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution

F1 Score = 2\*(Recall \* Precision) / (Recall + Precision)