Work report

This report is meant to illustrate the scope and flow of work in an industrial internship program at Haskayne School of Business under Dr. Marco Bijvenk in collaboration with MNP limited. The internship was with the Machine learning (ML) and artificial intelligence (AI) team of MNP and was supervised by their AI/ML team and Dr. Bijvenk. We worked with two AI/ML use cases namely in mergers and acquisition and retail marketing client conversion. We tried to record all the shortfalls and learnings as they were realized, to relay them to be used in the future.

The report includes problem definition of AI/ML use cases, exploratory data analysis, and AI/ML modeling methodologies serving the business problem. The programming environment used is python. Outcomes are validated using standard methods and results are good to be used for similar use case scenarios.

Project 1

Mergers and Acquisitions (M&A): Do the required market research and create a dataset to be used to predict the likelihood of M&A between companies A and B.

Understanding the ML use case data requirement in the M&A landscape:
Our target here is to understand the M&A activity as a process and collect data required to represent all key indicators in a successful or failed M&A. The dataset should be robust to represent the following:

1. Type of M&A:
   a. For size growth
   b. To proactively capitalize on current and future market growth in a particular business. For example, Salesforce acquires Slack.
   c. For product or functional synergetic M&A.
   d. Predatory acquisitions
   e. For tax savings

The above list is not exhaustive.

2. Data on key performance indicators (KPIs) of M&A [1]
   a. Financial data. (revenue, taxes, market capital, assets, debts, costs etc.)
   b. Customer KPI
   c. Process KPI
   d. Employee KPI

3. Apart from the above, data should also talk about the success or failure of the M&A. This information will be the basis of classification {success:1, failure:-1}.

Historic financial data of completed acquisitions are hard to find. To overcome this limitation, we looked for M&As announced or in process in the years 2020-21. Data can be downloaded from Yahoo finance using freely available python packages. The script for data download from Yahoo is appended in the appendix.

Compustat has historic data of many companies acquired in past. It is easy to download the data from Compustat in desired file format. Downloaded and cleaned data for missing and repeated values are included in Appendix. An account is needed to download data.

At this stage, I realized we need to do a lot of text processing on various reports to get other KPIs which demands time and resource beyond the internship goals. The appendix includes all the codes pertaining to the data download task.

Project 2

Given data of direct marketing campaign, exploit data-driven analysis and ML modeling to serve the following retail business objectives:

- ***Segmentation and targeting of clients.***
- ***Expense optimization for future campaigns.***
- ***Sales forecasting.***

Retail marketing has been myopic in the absence of an accurate sales forecast. Reports suggest only 30% of sales forecast is predicted accurately 90 days out [2]. One of the main reasons behind this setback is dynamic demand patterns. In today's data-rich market, data-driven marketing is being widely accepted but still, we record failure in bridging the demand-sales gap. Another article [3] suggests that 74% of companies are becoming more data-driven but only 29% connect analytics to action. These facts and figures necessitate a platform where analytics get more meaning with data. We will exploit AI/ML to provide such a platform where retail businesses can be more confident in understanding their business, client, and forecasts.

Direct marketing campaigns are used in retail marketing to connect to clients and collect the responses to their products and services. Retailers reach out to the clients through different modes of media and impress on them certain advertisements or calls. Such campaigns are scalable depending on the performance of previous similar campaigns. Also, these campaigns help collect data on their client base and provide a solid gauge on the outcomes of the campaign. We can think of embedding AI/ML in such direct marketing campaigns as shown in Figure 1.

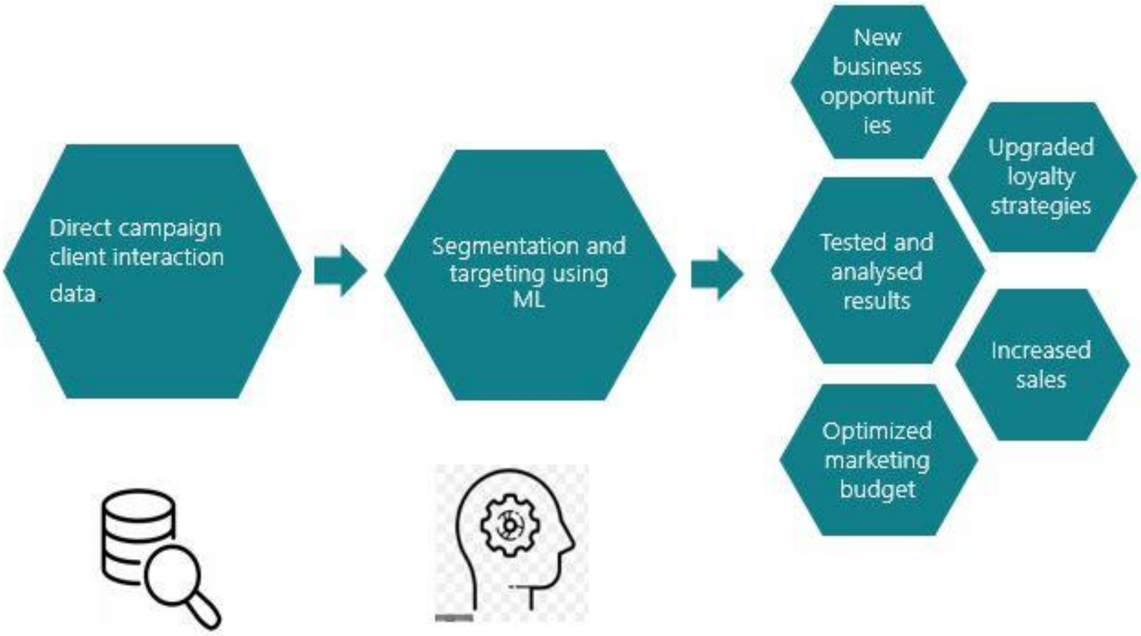
Direct campaign client interaction data.
Segmentation and targeting using ML
New business opportunit ies
Upgraded loyalty strategies
Tested and analysed results
Increased sales
Optimized marketing budget

Figure 1. AI/ML embedding in direct marketing campaign objectives

A possible AI/ML use case in such a marketing campaign is discussed in the following sections:

- The problem
- The data
- Exploratory data analysis
- ML modeling and predictions
- Conclusion and future work

The problem

A company xyz has collected data on their previous direct marketing campaigns through social media platforms. The company's interest is to use data-driven market analysis and forecast to serve the following three goals:

- Segmentation and targeting of clients.
- Expense optimization for future campaigns.
- Sales forecasting.

An AI/ML-based pipeline can be a comprehensive solution to all the above-mentioned business objectives. In this report we will be majorly looking at the problem from a development perspective. We will also briefly discuss ML pipeline deployment for production as a front-end dashboard or any other visualization mode.

The data

Quality and structured data are central to any ML-based analysis. Available data [data link] in the present case is provided by the company xyz which includes 1143 datapoints with the following 11 features.

- ad_id: a unique ID for each advertisement.
- xyzcampaignid: an ID associated with each advertisement campaign of XYZ company.
- fbcampaignid: an ID associated with how Facebook tracks each campaign.
- age: age of the person to whom the advertisement is shown.
- gender: gender of the person to whom the ad is shown.
- interest: a code specifying the category to which the person's interest belongs (interests are as mentioned in the person's Facebook public profile).
- Impressions: the number of times the advertisement was shown.
- Clicks: number of clicks on for that advertisement.
- Spent: Amount paid by company xyz to Facebook, to show that advertisement.
- Total conversion: Total number of people who enquired about the product after seeing the advertisement.
- Approved conversion: Total number of people who bought the product after seeing the advertisement.

In the further discussion, we will be referring to these feature definitions. Each sample or data point tells the response of showing a unique advertisement to a particular age group and gender with some Interest domain. The responses are captured as clicks, Total conversion, and Approved conversion.

Exploratory data analysis

In many cases, data obtained is incomplete or missing. The data type is non-uniform with strings, numeric, date-time, Boolean etc. We need to ensure consistency in data and make it accessible for manipulation. In the present case, the data is complete with no missing values, hence we saved this stage. We are using pandas data frame as a labeled data structure for accessing the data.
At this stage, we will map the objectives to the set of queries we need to ask for the data.

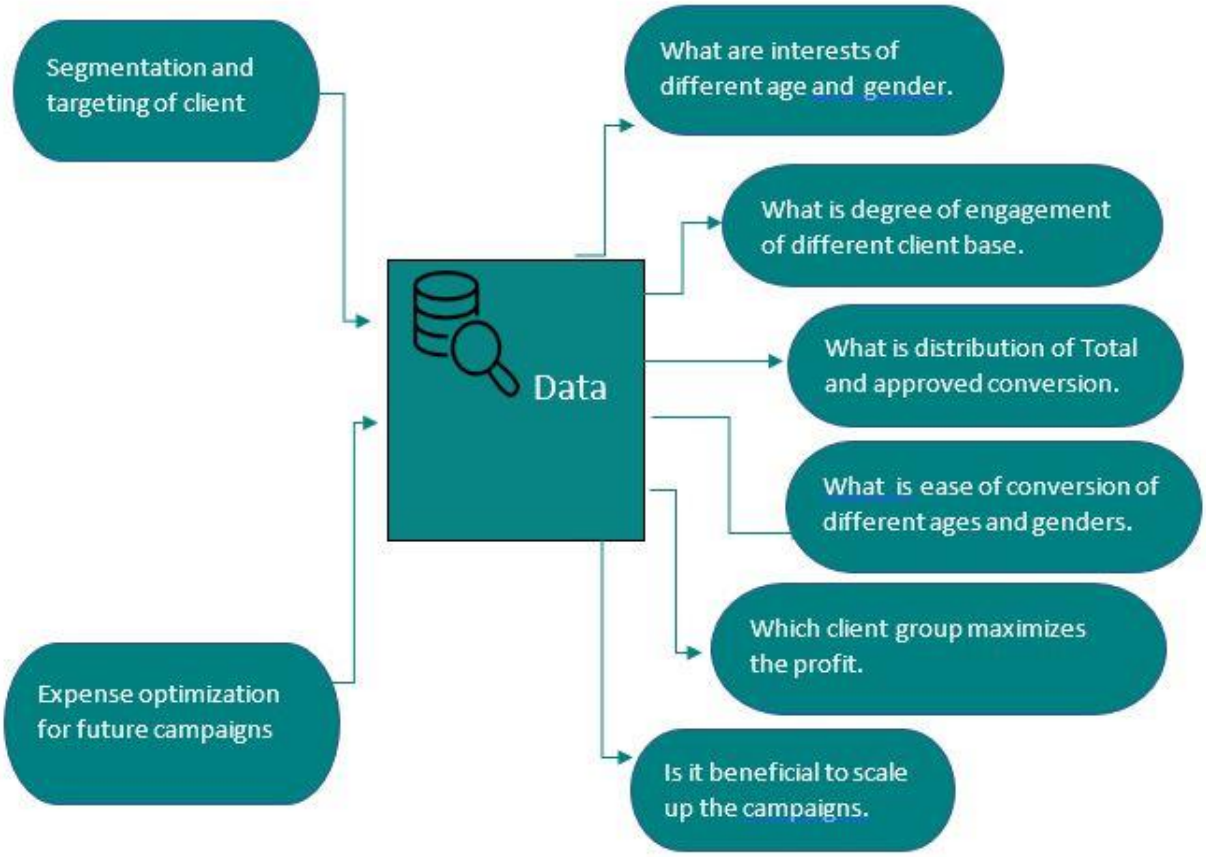
Segmentation and targeting of client
Expense optimization for future campaigns
Data
What are interests of different age and gender.
What is degree of engagement of different client base.
What is distribution of Total and approved conversion.
What is ease of conversion of different ages and genders.
Which client group maximizes the profit.
Is it beneficial to scale up the campaigns.

Figure 2. Objective mapping of business interest to data queries

We exploit data visualization capability of pandas dataframe along with seaborn and matplotlib to respond to the queries mentioned above. We will look at scatter plots, histograms and box plots to draw useful conclusions

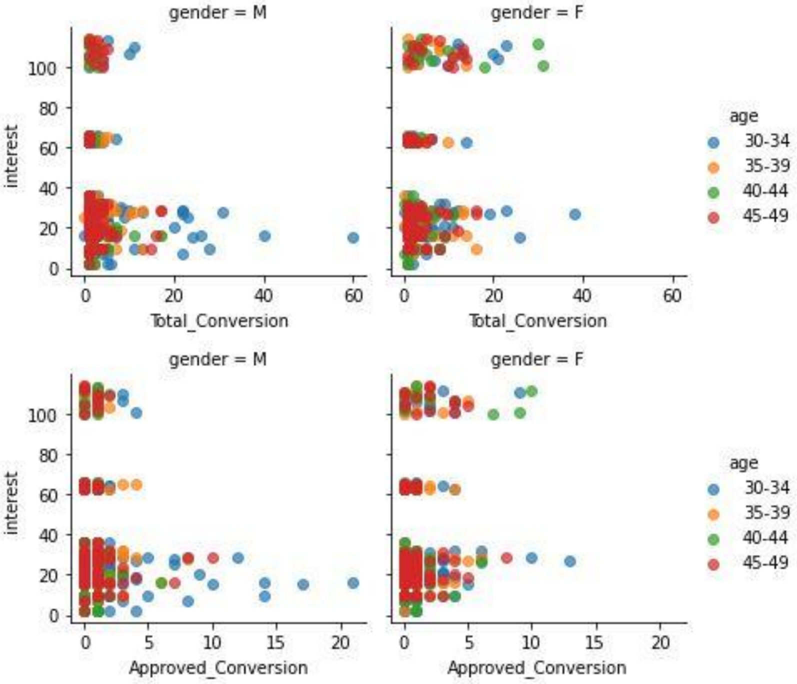

Figure 3. Scatter plots showing distribution of Approved conversion with Interests

Scatter plot indicates that major Interests are within 0-40 and majority of high number of approvals came from this interest group.

More females than males have Interest in 100 and beyond and so is the conversion.

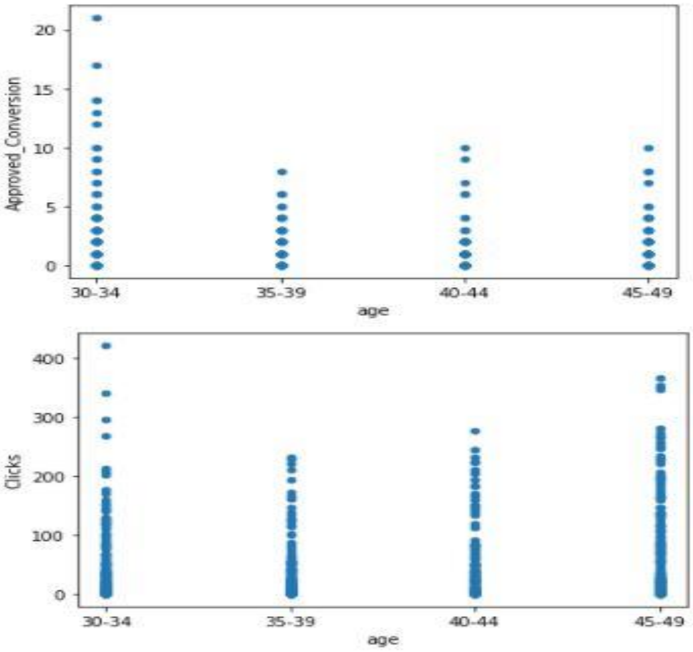

Figure 4. Scatter plots showing distribution of Approved conversion and clicks with age.

Approved conversion and clicks ratio are worse in age group 45-49 [Figure 4].

Approved conversions are maximum in age group 30-34 males [Figure 3 and 4].

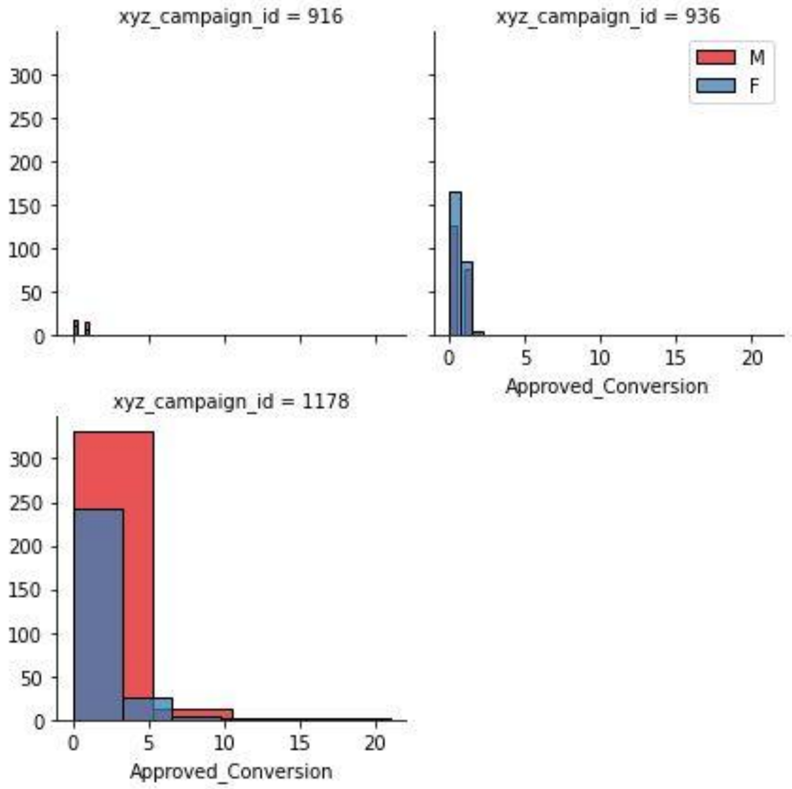

Figure 5. Histogram showing frequency distribution of Approved conversion in different campaigns.

This histogram shows approximately doubling the participation increased the Approved conversions more than five folds. This means campaigns are scalable and there is significant scope of participation enhancement.

Correlation matrix may provide further insight into the data.

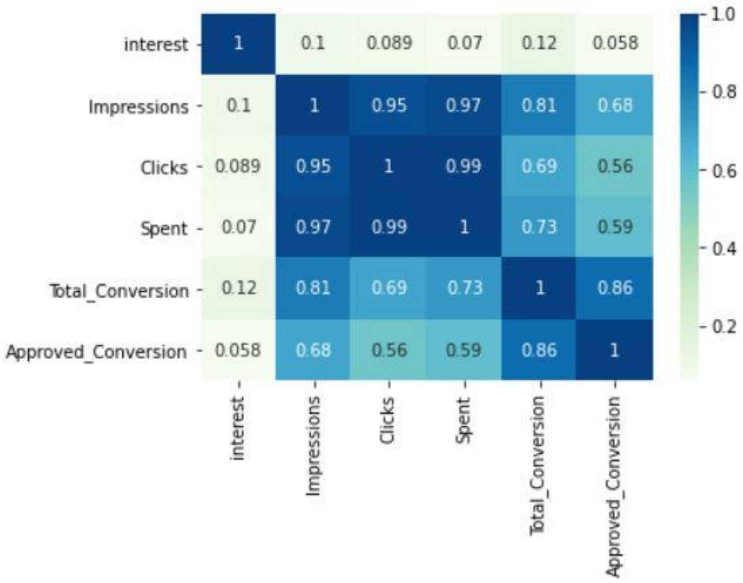

Figure 6. Correlation matrix of different feature

- Impression has strong correlation with all numeric variables which is obvious.
- Impressions, clicks and spent have very high correlations.
- There is 0.6 correlation of click and spent with approved conversion. Looks like our ads are putting some effort but not sufficient.
- There is some limited scope of improving on Impressions/Click relation. Which in turn will improve Total conversions and hence Approved conversions. Impression can improve the conversion.

With a better insight into the data and whole lot of information about relationships between features we can move on to the third objective which is predictive in nature. Here we will continue

ML modeling and predictions

ML embedding into a system essentially doing the following:

- Feature engineering
- Model selection
- Model validation
- Model extensions

Feature engineering

Refer figure below for the elements of feature engineering in this project.

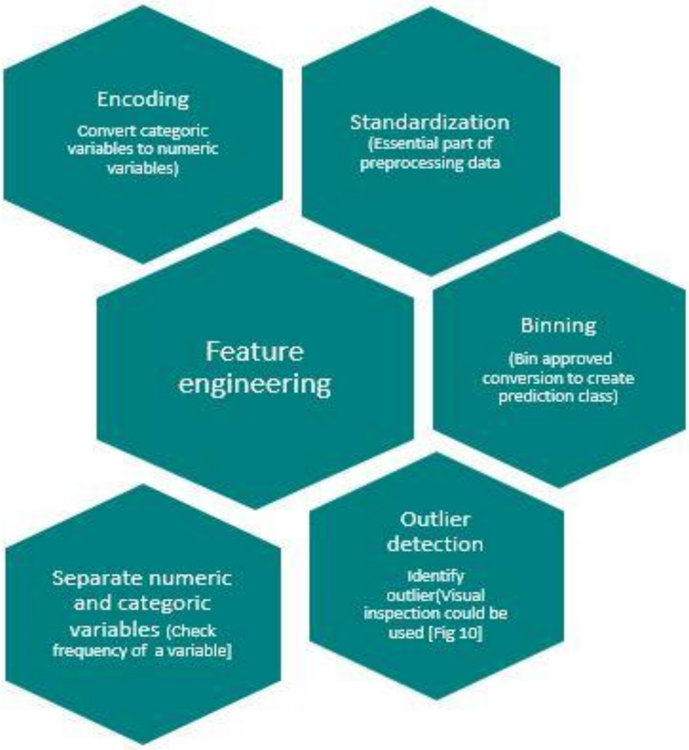
Encoding
Convert categoric variables to numeric variables)
Standardization
(Essential part of preprocessing data
Feature engineering
Binning
(Bin approved conversion to create prediction class)
Separate numeric and categoric variables (Check frequency of a variable]
Outlier detection
Identify outlier(Visual inspection could be used [Fig 10]

Figure 7. Feature engineering elements in this project

We create here new features depending on prediction requirements. Standardization is the scaled distance of features from its mean. It is essential to do this step as part of preprocessing as our data values are varying in a wide range..

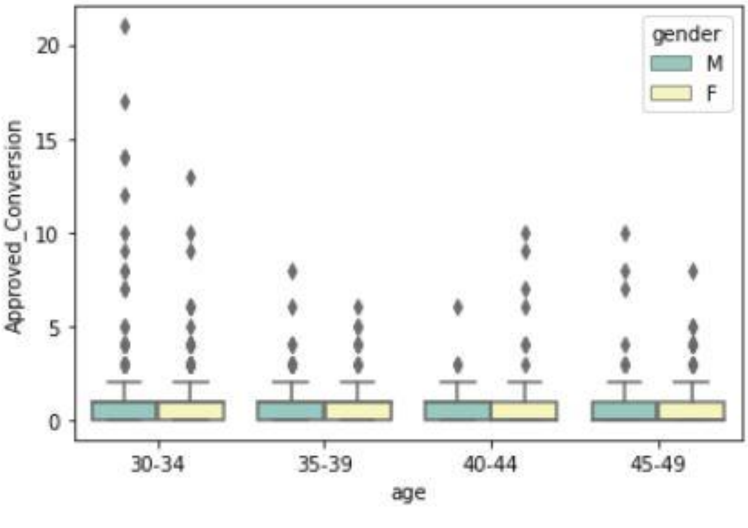

Figure 8. Showing dots as outliers and distribution of approved conversion with age.

Model selection

The classic confusion of 'regression or classification' is not a problem here as we are trying to classify a data point based on its conversion status or clicks. In our case, we are focusing on conversion status.

Here a support vector classification approach will be suitable. We need to add a feature which is predicted class. For that, we can group the Approved conversion in 4 bins and create a new class out of it. Alternatively, we can cluster the data in an optimal number of clusters and give those cluster labels as predicted class to SVM classification.

Following is benchmarking of results from the different methods of classification:

Method	F1_score	Time of execution (s)
SVM svc only	0.945	0.76
K mean and SVM svc	0.99	0.3
KNN	0.78	0.3

The optimal number of clusters for K means was obtained by the elbow method, which becomes crucial if we are going to use those cluster labels for further classification.

Model Validation

Although the data is split in training and test data and we check the model performance with test data set, but there may be still biases in the test data selection. To overcome this issue, we used cross validation method which essentially splits the data into *k* different train test sets. Train data is fitted on the estimator and accuracy scores on the test set are calculated iteratively for each such split. Finally, we average over the scores to see model performance. This method of cross validation eliminated the requirement of separate validation dataset.

Model extension

Serving to the goal of ML embedding in the marketing predictions, we need to develop the model for large scale uses. Keeping this idea of scalability in mind I introduced pipeline method in the model which arranges steps to be followed (preprocessing and classification with cross validation) in a sequential manner. A schematic production diagram is shown below.

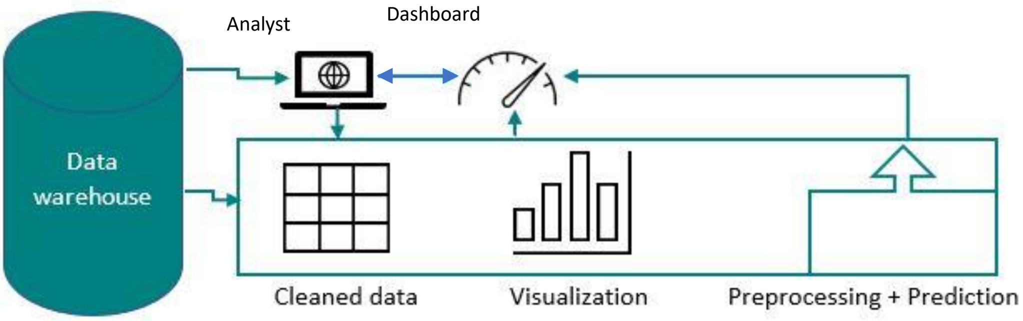

Figure9. Schematic diagram of a ML pipeline for production.

Conclusions and future work

1. Improving on Impression will certainly improve the approved conversions, however, in some age groups, it is observed that the clicks are not getting converted which means either the ad is not communicating, or the product needs to be customized for that age group.
2. The gap between total conversion and approved conversion needs to be filled. We can try NLP of phone conversation records for this purpose.
3. There are certain Interest areas liked more by women than men. Quality of ads may play important role there to improve clicks and hence the conversions.
4. In general women click on ads more and get converted less than men, except in age group 40-44.
5. SVM classification with K-means clustering is very accurately predicting the client conversion. Model testing results are good. This method has advantages over other classification approaches in response time and quality of prediction.
6. The Pipeline approach can be used for scaling up the model to production level.

References:

[1] M&A KPIs: The Right Merger and Acquisition KPIs and Metrics to Prevent Failure, Rhythm systems. Dec 1 2020.

[2] Gabe Larsen, Stephen Hurrell and A J hunt 'The Gap Between Sales Forecast and Reality' InsideSales Labs CSO report. Feb. 2019

[3] Hopkins B**.** **'**Think You Want To Be "Data-Driven"? Insight Is The New Data', VP, Principal Analyst, MAR 9 2016

[4] Scikit documentation

[5] Trivedi S., Zachary A. Pardos and Neil T. Heffernan, 'The Utility of Clustering in Prediction Tasks', Department of Education IES Math Centre for Mathematics and Cognition grant. Report Date: 05 September 2011.

[6] 'Process Manual From ML to M&A:, Ten M&A Target Predictions through a Machine Learning Model'. Wisconsin school of business December 16, 2019.