Arbitrium

Documentation

Contents

Introduction

Arbitrium, Latin for Decision, is a business dashboard built to ease the decision making process. It aims at providing insight to DHL regarding the factors involved in client retention. It works by predicting the churn rate and the probability to win a contract with a customer. It also provides categorical company wise data, meant to provide further insights.

The product has been developed taking into the following factors:

- Useability
- Scalability
- Maintainability
- Simplicity

We will provide the details of each the modules present in our product.

Module 1: Client Performance

The aim of this module is to provide consolidated performance insight using both publicly available data and the data provided by DHL for company performance evaluation.

We have used a scoring model termed as "Client Score" in order to evaluate each client. The model takes into account various parameters such as the stock market performance over the last 5 years and the public perception towards the company.

Equity Performance: A major reason the company share price is often a valuable metric is that is used an indication of the overall strength and health of a company. A high growth of the company is often directly reflected in the share price and is thus an extremely valuable metric.

Social Media Perception: To provide a holistic view of the company's performance, we have utilised libraries such as the VADER Sentiment Package to analyse the target company's customer perception.

Working

The customer score index is a score generated from the weighted average of social media analytics and the stock-market performance.

Social Media Analytics: Data is streamed from Twitter constantly based ona set of very specific keywords for each company. The data stored is a mLab database (mongoDB). The tweets are collected from the database and the sentiment of each individual text is calculated and a single average score is returned for a company.

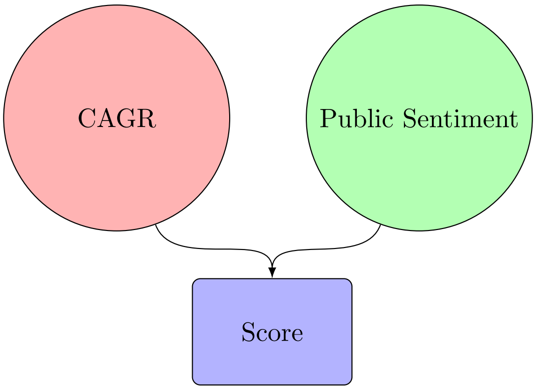
The data for the sentiment modeule were posts from Twitter that mainly consisted of short text. After analysing multiple models, we found that the VADER Sentiment Analysis lexicon based model provided the best performance for this data.

Advantages

- Support for text such as: "The company has not been performing well". A regular word based model would give it a positive score as most models neglect the presence of the word 'not'. However, the VADER Sentiment Module has support for use-cases like these, and thus provides a better analysis of the social media data.
- Support for emoticons and booster words: The module supports the emoticons widely used in social media. For example, ':)' tends towards a positive sentiment whereas, ':(' tends towards a negative sentiment. Moreover, words like 'very' and 'a lot of' tend to increase and emphasize the emotion and thus the final sentiment is boosted without changing the sentiment type (i.e. positive becomes more positive and negative becomes more negative).

CAGR: Compounded Annual Growth rate or CAGR is one of the most important metrics while judging a company's performance and growth. Having acquired the stock market data for every client, the CAGR is calculated every year.

The CAGR score coupled with the social media analytics score gives a customer index which can be used by DHL to analyse the performance of the client. A poor performing company would have a high change of leaving DHL as a partner, due to possible cost-cutting measures. The client may also be closing down its offices or warehouses in particular locations. This eliminates delivery requirements from that location.



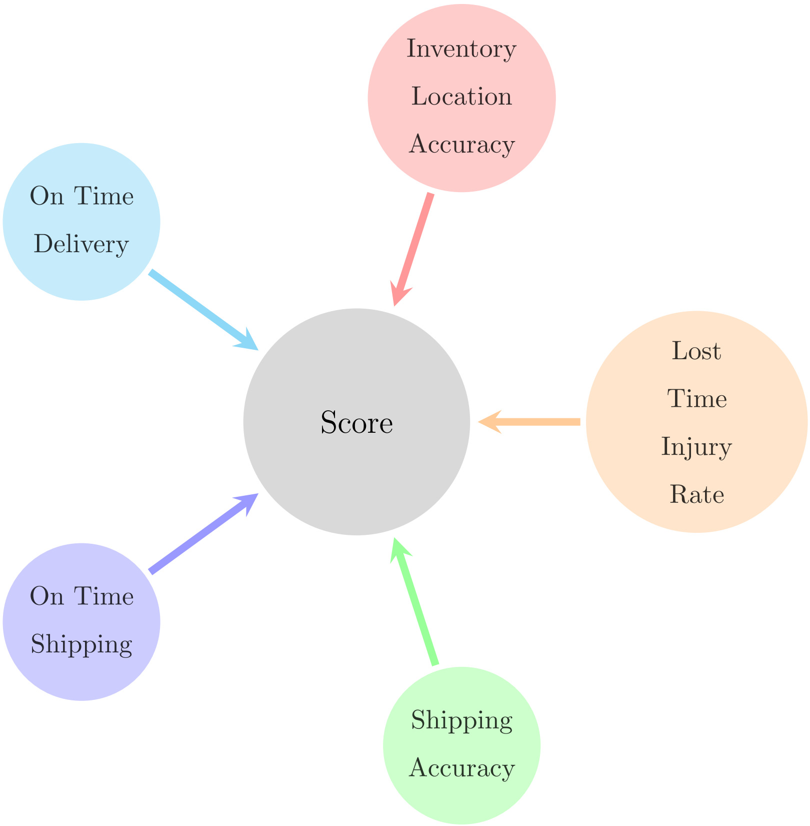
Module 2: DHL - Target Company Performance

This module is aimed at providing a country wise analysis of the Key Performance Indicators (KPIs) score. Each of the scores take into account the following indicators:

- Inventory Location
- Lost Time Injury Frequency Rate
- On Time Shipping
- Shipping Accuracy
- On Time Delivery

The scores provided above are important, while taking into account the following considerations:

- Delays - determined by Lost Time Injury Frequency Rate and On Time Shipping. A lower delay is preferred. This helps in reducing the soft costs.
- Synergy between the companies - determined by Inventory Location Accuracy and On Time Delivery.
- Higher Costs Incurred - determined by a combination of all the above factors.



Working

In order to calculate DHLs performance with respect to a particular client, a weighted score is derived based on the parameters mentioned above. Lost Time Injury Frequency Rate has an negative impact on the performance while the others have a positive impact on the performance. The final score has been calculated as:

$$score = 0.2 \times ILA + 0.2 \times OTD + 0.2 \times OTS + 0.2 \times SA + 0.2 \times (100 - LTIFR) \qquad (1)$$

$$\begin{aligned} where, ILA &= InventoryLocationAccuracy \\ OTD &= OnTimeDelivery \\ OTS &= OnTimeShipping \\ SA &= ShippingAccuracy \\ LTIFR &= LostTimeInjuryFreqencyRate \end{aligned}$$

This score calculated gives a measure of DHL's performance towards the client. Clients for whom this score is low, would be looking at other alternatives to DHL and this would lead to DHL eventually losing revenue from the client.

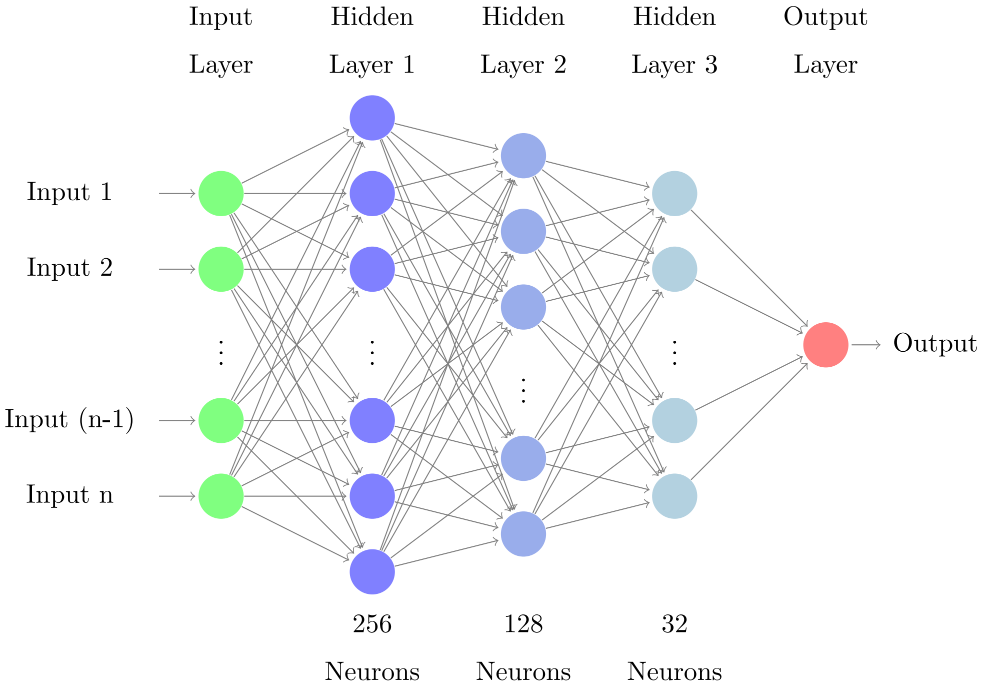
Module 3: Predictive Analysis

This module is aimed at predicting the probability of DHL winning the target company as a client, and by extension, the churn rate.

The primary prediction by this model is the probability to win. In other words, it predicts the chances that a client successfully completes a contract. This metric depends on a variety of myriad factors. Current research in this field indicates that Decision Trees and Neural Networks have the best performance out of the various classifiers such as Naïve Bayes, Support Vector Machines and Logistic Regression. Thus, we have used the combination of a Neural Network and a Random Forest as our predictive model.

Neural Network

The Neural Network is used to recognise general patters between certain factors and the probability to win a client. The architecture used is a multi-layer or a Deep Neural Network, as shown below:



Neural Networks are very sensitive to the quality of data which is fed to it. If the data isn't pre-preprocessed, it becomes very difficult for the net to settle at a good value of weights. As a result, the data was pre-processed using the following steps:

- Data Cleaning
  - Missing values were replaced with 0. The reason for this is that 0s do not affect the performance of the Neural Net as they do not trigger any activation in the neuron.
  - NaNs were converted to zeros based on the previous reasoning.

  - Unnecessary features were removed, based on preliminary exploratory data analysis.

- Data Manipulation
  - In a case like this, data such as dates have little significance on their own. They were converted to reflect the duration taken to reach a particular milestone.
  - The dataset provided had a lot of categorical data. This cannot be directly fed to a Neural Net. This was mapped to a numeric value using a fuzzy mapping algorithm.
  - The values contained in the data had a large variance. It was normalised to lie in the range -1 to 1. This increase the contrast in the Neural Net.

The data was split into a training set and a testing set, with a 3:1 ratio respectively. The weights were randomly initialised using a normal distribution having a standard deviation of 0.1. The Tanh and Sigmoidal activation functions were used. The Adam Optimizer was used with an exponentially decaying learning rate. Accuracy was evaluated on the testing set, with a 10% variation permitted. Using this architecture and parameters we got an accuracy of 85%.

Random Forest

The purpose of using a Random Forest was to predict the probability to win. For each company that DHL has, a separate classifier is trained. This ensures that the prediction is unique to a company. In this model, the categorical data was handled in two ways:

- Mapping: The categorical features were converted to numeric values using the same fuzzy mapping algorithm used in the Neural Net.
- Encoding: Categoricl features were split into columns. Each column was mapped to a specific category of the feature. The columns were then populated with 1s and 0s. A Random Forest Regression does not depend on the actual magnitudes of the values present. Hence the choice of 1s and 0s were trivial.

Each classifier for a company was trained on the data specific to that particular company. 200 decision trees were used to train the model.