Problem 1: Churn (Customer & Employees)

Churn is when a customer or employee leaves the company. Churn is a very costly problem that eats away at the organization's revenue when customers leave or employees quit.

Churn is a **binary classification problem** where we are trying to determine whether or not something will happen. Hence either a 1 or a 0. However, the 1 or 0 is actually just a label that is based on a class probability (e.g. class 1 = 0.75 and class 0 = 0.25).

To solve this problem, I use:

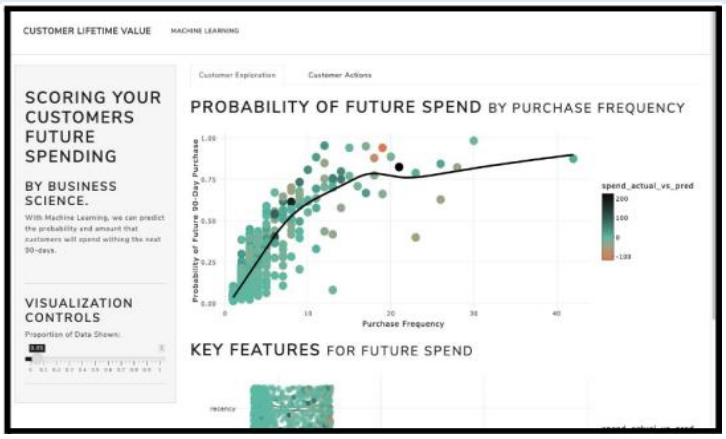
- **H2O Automatic Machine Learning** to predict the churn risk of each employee
- **Local Interpretable Model-Agnostic Explanations (*LIME*)** to explain why the model predicts the employee as likely to stay or leave the company.

Problem 2: Customer Lifetime Value (CLV & RFM)

Customer Lifetime Value (CLV) is used to highlight which customers to focus on. Companies use CLV to **estimate the profitability** of the future relationship with a customer. Most algorithms use **RFM (Recency-Frequency-Monetary)** attributes for customers to help estimate customer lifetime value.

There are two approaches to CLV:

1. Traditional economic
2. Machine learning



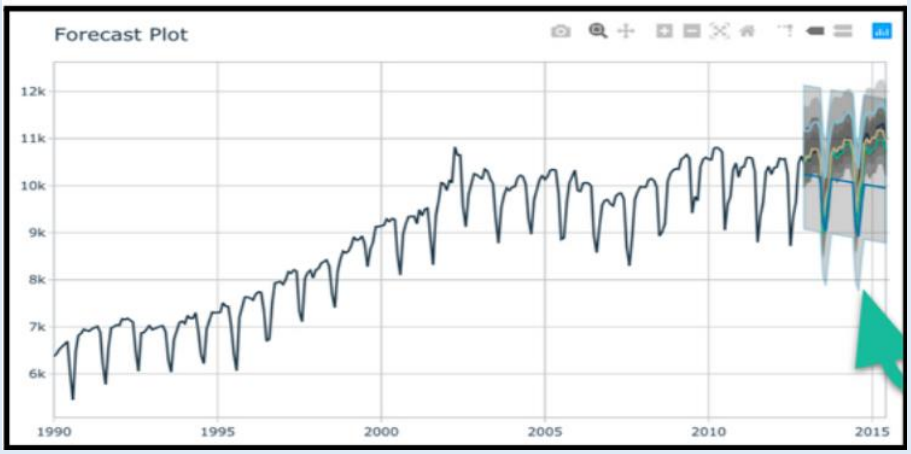
I favour machine learning as I normally get higher accuracy and better predictive insights when I model using the newer technology (machine learning).

Companies use this metric to gauge profitability and **focus on which customers**. In short, CLV is the **profit** estimated by the future relationship with a customer. There are **many different approaches** to modelling CLV.

The customer lifetime value app **scores the customers** by estimated future purchasing in the next 90-days.

Problem 3: Time Series Forecasting

Time series forecasting is the process of using historical data (such as sales revenue over time) as inputs to predict (estimate) the next H periods (often denoted the forecast horizon). Everything from hiring new employees and purchasing raw materials via the supply chain is driven off of a demand forecast. Therefore, organizations that improve their forecasting can save millions of dollars.



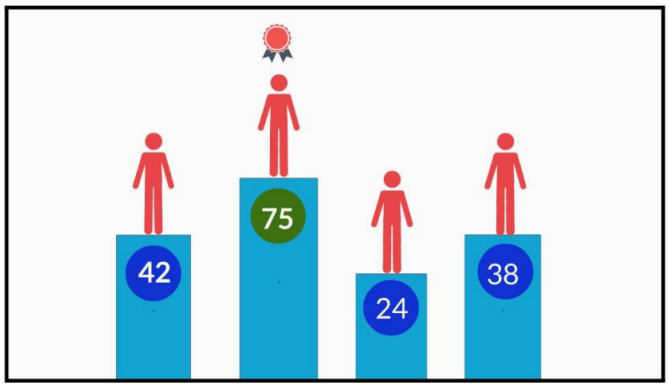
A demand forecast may look something like above for a *single time series.*

Companies need to make **hundreds of thousands of forecasts** every week or month (think of forecasting for every product your company makes). This is called **high-performance forecasting**.

Problem 4: Lead Scoring (Email & Sales Marketing)

Lead scoring is a binary classification problem where you want to predict the probability of a lead turning into an order. Traditionally this is done by measuring the number of actions taken by a customer and then calculating a score by weighting each action in terms of the business's idea of how important the action is.

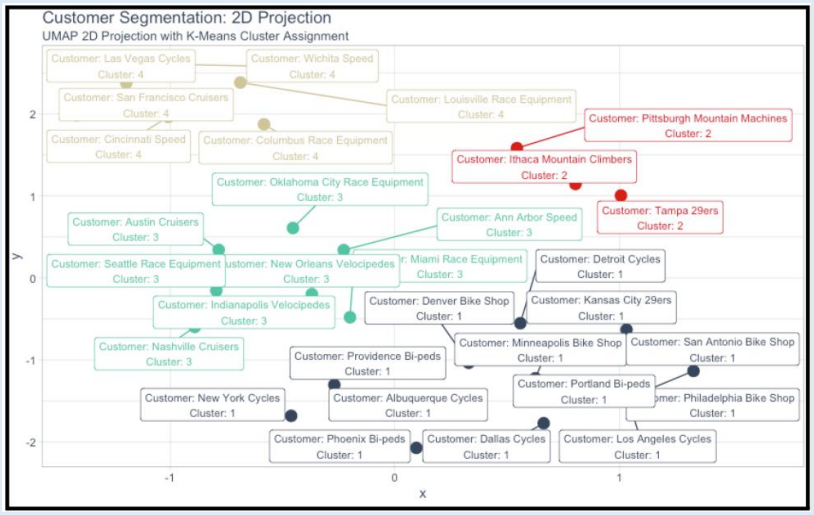
Data science can be used to more accurately predict the customer's likelihood by combining the customer's actions taken (traditional method) and additional data like customer's time with the company, age, geographic region, and more. The result is a more accurate lead score.



Lead scoring is a variant of Problem 1: Churn. The main difference is that instead of classifying the customer as Yes/No, we wish to **rank each of the customers** in order of most likely to purchase to least likely. Customers that are more ready to purchase can be marketed to, and those that are least likely should be nurtured.

Problem 5: Customer Segmentation

Customer segmentation is the process of separating customers into groups based on common purchasing behaviour (e.g. this group tends to buy more shoes vs this group tends to buy more handbags). Companies can then target these segments and tailor marketing messages most commonly via email. **Segmentation increases revenue and improves customer satisfaction. Win-win.**



You use clustering and dimensionality reduction to visualize segments of customers using:

- **K-Means:** A common clustering algorithm that does well at identifying segments within business data
- **UMAP/PCA:** A dimensionality reduction technique that allows us to plot the variance of the data in two dimensions

The result is a customer segmentation that can be used to target similar customers with tailored marketing messages.

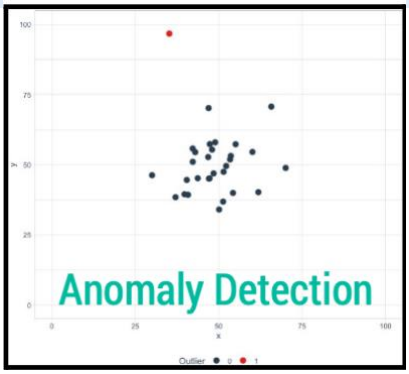
Problem 6: Fraud Detection with Anomaly Identification

Fraud detection is identifying abnormal transactions or behaviour from customers. Often the abnormal behaviour can indicate that something isn't right about the transaction or insurance claim, and therefore requires a hold until further evaluation can take place.

Key issues:

What does **abnormal behaviour** mean?

Which techniques work gracefully in the presence of **High Imbalance**?



A common method is **anomaly detection** which is used to identify abnormal events. This can be interpreted as suspicious behaviour.

Problem 7: A/B Testing (Website & Emails)

A/B testing is the process of testing changes to websites and emails by dividing the traffic into two and randomly assigning to two variants. It's most common to make small changes to each variant, and then to test the impact using statistical analysis.

Problem 8: Production & Deployment

Production is the process of taking a machine learning algorithm or decision-making analysis code (**script**) and converting it into a form that is usable by the business. For Python, the most common methods are to build and deploy web apps using **Dash** or **Streamlit**. The most common production app is a dashboard. Think of the app as the package for the code, and users interact with the app (***using dropdowns and buttons***), which then automates the process of running the code behind the scenes. (***ChatGPT can be used for this problem***)