Cluster Based Forecasting

Nicholas Santoso

1. Problem

Currently Salesforce Forecasting allows users to forecast opportunities by user, product family, forecast category, ect. The limitation with Salesforce Forecasting is that it can only aggregate opportunities by the fixed dimensions listed above.

In many cases it is much more meaningful to group opportunities that share similar characteristics, and monitor forecast numbers of each group. This is because similar opportunities tend to have similar trends and behaviors, and groups of similar opportunities tend to have independent trends and behaviors from each other. The term that usually describes this process of splitting opportunities into groups of similar opportunities is called *segmentation*.

In addition to the benefits of more insightful forecasts, segmentation of opportunities can potentially improve the overall forecasting prediction accuracy. This is because instead of creating one prediction model for all of the opportunities, we can create an independent model for each segment of opportunities.

But before users can start tracking forecast numbers of segmented opportunities, there needs to be a tool to help identify a useful and informative segmentation.

2. Solution

This hack day project aims to provide an end-to-end solution by first providing an interactive tool that users can use to visualize and identify a useful and informative segmentation in their opportunity record data. Second, it automatically aggregates your opportunities by the segments the user has identified with the tool. Third, it potentially improves einstein forecasting predictions by generating independent models for each cluster of opportunities.

2.1. Opportunity Clustering

The first step is to identify how to group opportunities by their similarity. We first import opportunity record data from Salesforce and plot it for the user to see in the UI.
To split the opportunities into groups of similar opportunities, we use a clustering algorithm called k-means. We decided to use a clustering algorithm for this because opportunities that cluster together often share similar characteristics and behaviors and tend to generalize well together. We first let the user specify which 2 fields they would like to cluster over, for example, OpportunityAmount and ProductType. We then plot all of the opportunities where the user can visually identify any clusters. Once the user has visually identified clusters in the plot, they can then specify the number of clusters they would like the clustering algorithm to find (k-value). An essential step before we run the clustering algorithm, is we normalize the data to ensure the variance between each dimension is the same. We then run the clustering algorithm in real-time, which labels each opportunity to a particular group, hence defining a segmentation/grouping.

It is beneficial to let the user interactively set the dimensions and number of clusters (k-value) because they can take advantage of their domain knowledge to determine which dimensions to cluster over. The other benefit of making the clustering process interactive is that the plot and clustering analysis might bring new insights into the data that the user might not have otherwise discovered.

2.2. Cluster Based Forecasts

After successfully identifying individual clusters in the opportunity data, we can now aggregate all the opportunities for each cluster and plot the forecast of each cluster. We then periodically run the clustering algorithm and aggregate to provide the most up-to-date forecast numbers for each cluster. We could potentially save the computational time of running the clustering algorithm by creating a voronoi graph, which statically defines which cluster an opportunity belongs to by the value of its parameters. The issue with this approach is we assume that the centroid/mean of the clusters do not move over time. In practical applications, this is not usually the case so it is best we dynamically re-cluster every opportunity on every update.

2.3. Improved Predictive Forecasting

The final step is to take advantage of the clustering to improve the overall forecasting prediction accuracy by generating a unique predictive model for each cluster of opportunities. We then plot the projected forecast prediction for each cluster. We also aggregate all of the predictions of all of the clusters to show the improved overall forecasting prediction for all of the opportunities.

The current implementation of predictive forecasting attempts to generate a predictive model for all of the opportunities. The issue with this approach is that the predictive model might require a higher complexity if there are multiple groups of opportunities that have independent trends and behaviors. By identifying independent groups of opportunities (clusters) and creating a predictive model for each cluster, we are able to reduce the overall complexity of the model and are able to deliver more meaningful insights on how each cluster contributes towards your predictive forecast as a whole.

3. Future Features

Further iterations of this system can allow for n-dimensional clustering, where the user can select a variable number of dimensions to cluster their opportunities over. There are also existing tools to help visualize high-dimensional data like t-SNE. We can also utilize other variations of clustering algorithms like categorical clustering, which is more suitable for clustering over dimensions with nominal variables. We can also use an online-clustering algorithm that can incrementally categorize new and incoming opportunities.

4. Related Existing Solutions

While there are several business intelligence tools that can help identify clusters in your data (like Tableau), these tools do not provide an end-to-end system that allows the user to interactively see the impact the clustering of opportunities have on their forecasts and predictive models. Another difference is this system is designed to be automated. Users can monitor the real-time forecast numbers and the predictive model performance of each cluster.

5. Demo

Link:
http://nicholass-wsm2.internal.salesforce.com:8000/
I would like to see if I can cluster my opportunities by business-size and product-price.
I might see 2 clusters:

1. A cluster of opportunities from smaller businesses interested in the essentials version of our product.
2. A cluster of opportunities from larger business interested in the enterprise version of our

product.

With cluster based forecasting we can find the aggregate of all opportunities in the 1st cluster and the forecast of opportunities in the 2nd cluster.

Instead of aggregating opportunities by simple categorical dimensions, with cluster based forecasting, we can aggregate by complex regions.

6. Technology Stack

* I am primarily using Plotly Dash as both the server as well as plotting tool, written in python.
* Scikit-learn is the library used for clustering