Learning from Route Plan Delivery for Last-Mile Delivery

By: Yiyao Li, William Phillips
Advisor: Matthias Winkenbach

Topic Areas: Database Analytics; Last Mile; Transportation

Summary: This capstone project studies route deviations in the last-mile deliveries of a large soft drink company. We study whether delivery crews systematically, consistently and substantially deviate from the planned stop sequence of their routes. Additionally, we analyze what drives these deviations and whether they add economic value or not. With this objective, we perform regression analysis and build classification models, using one-year data across two countries, Mexico and USA.

Before coming to MIT, Yiyao graduated from a B.Eng. in Engineering Systems Design from the Singapore University of Technology and Design. She then worked at Visa Inc, across Corporate Strategy, Product and Merchant Sales and Solutions departments.

Before coming to MIT, William worked in Walmart's logistics division for five years. Prior to that, he graduated with a Bachelor's degree in Engineering from Pontifical Catholic University of Chile. Upon graduation, William will join Amazon's retail area.

KEY INSIGHTS

1. Using environmental variables that describe the route, drivers' decision to deviate from the plan can be predicted with an accuracy of 71% in Mexico and 84% in the US.
2. With the same environmental variables, the impact of deviating from planned sequence on distance can be predicted with a coefficient of determination R^2 of 0.57.
3. Drivers are more likely to deviate and increase the route's distance when more customers are visited. Efforts should be focused on these routes.
4. Customers' geographical locations, reflected in the ZIP codes, are highly useful to predict deviations.

Introduction

Route deviation in last-mile delivery is a critical problem due to its substantial economic impact in operational cost. This capstone project is a quantitative analysis and modeling effort to assess the impact of sequence deviations of actual last-mile delivery operations from optimal route plans based on a soft drink company's data in Mexico and the USA.

Drivers stated preference in selecting routes has been extensively examined in the literature through surveys methodology. However, the recorded deviation of routes which are revealed preferences has not been studied extensively as drivers stated preferences have been studied. This capstone project is focused on filling this gap by assessing recorded one-year data on route deviation which are revealed deviation in multi-stop last-mile delivery.

We create metrics to measure deviations and set them as target variables to predict deviation. These metrics includes Deviation, Sequence Deviation, Distance Deviation, and Straight-Line Distance Deviation. Based on the preliminary assessment of the dataset, we observe significant differences between the USA and Mexico, in terms of data size, routes' characteristics, and deviation behavior. So, we recognize the geographical difference and separate the modeling for the USA and Mexico. Next, we model the Deviation, and Straight-Line Distance Deviation variables using three methods, namely Linear Regression, Neural Network and Random Forest.

Data & Methodology

The data used in this project comes from a worldwide beverage company with intensive last-mile delivery operations. It's divided in two tables: *Routes_Table* contains one row per route instance, representing particular route made on a given day. It includes time-related parameters (year, month, day, weekday), the distribution center (DC) from where the route was done, the number of customers served, and the actual and planned times and distances. *Deliveries_Table* contains one row per stop and includes geographical information of the customer and the drop size. It also contains comparison between what was planned and actually happened, such as the sequence position within the route instance, the arrival time and the stop duration. Lastly, by using the location of each customer and Google's API, we included we included ZIP code in the data base.

Regarding our data cleaning process, we only used route instances for which we had the information of both the route and of the specific stops. Secondly, we discarded route instances with empty fields that would afterwards be used in the analysis. Lastly, we performed basic analysis on the data to check its accurate. We found segments of the data (particular DCs and months) that had no deviations registered, and these were filtered as they were probably not accurately inputted.

This project is focused on the sequence deviation. To measure how much a route is deviated and the impact on distance that deviations have, we created four metrics. *Deviation* is a binary variable that takes the value of 1 when a route deviate and 0 otherwise. *Sequence Deviation* measures the fraction of the segments that deviated from the planned route sequence. *Distance Deviation* measures the percentage difference between the recorded and the planned route. This metric has the defect of taking into account the specific streets followed by the driver, which is outside the scope of the project. To isolate the effect that sequence deviation has on distance, we created *SLD_Deviation*. This metric measures the percentage difference between the distance of the actual sequence and the planned sequence, calculated using the straight-line distances (SLD) between the stops.

Figure 1 shows an overview of the data and deviations metrics. It stands out that 75% of routes deviate, however the impact on *SLD_Deviation* is only 2.6%. We see significate differences between both countries, in the number of observations and in the deviation metrics. We build a common framework of analysis, however because of the differences between both countries, it was applied independently to each one.

Figure 1: Data Overview Including Deviation Metrics

	Mexico	US	All
Route Instances	7,644	47,881	55,525
Number of DCs	9	9	18
Stops per Route	17.9	12.8	13.5
Route Distance (km)	73.0	106.9	102.2
Deviation	45.8%	79.8%	75.1%
Sequence Deviation*	61.8%	54.7%	55.3%
SLD_Deviation*	12.1%	1.7%	2.6%

** Only considering deviated routes*

Our framework has two models. One predicts the binary variable *Deviation*, and focuses on understanding what causes deviations. The second predicts *SLD_Deviation*, and focuses on predicting the impact in distance of the deviated routes. Each model was studied using regression analysis and classification methods.

Regression analysis gave us insights on the relationship and significance of each independent variable. Through an iterative process we eliminate the less significant variables. We determine the performance of our models by measuring the proportion of the deviations that each model explains, represented by sets of the coefficient of determination. We use adjusted R^2 for linear regression and generalized R^2 for logistic regression.

Once the independent variables are chosen, we build two classification methods: Neural Network and Random Forest. In general, these methods have the advantage of higher predictive capabilities, at the cost of little or no interpretability. We partition the data in two fixed sets, 70% as training set used to build the model and the remaining 30% as validation set, used to measure the predictive capability of the models. We also use generalized R^2 to as a performance metric. For the model predicting *Deviation* we additionally calculate accuracy, sensitivity and specificity. These metrics are calculated by classifying the results in a confusion matrix.

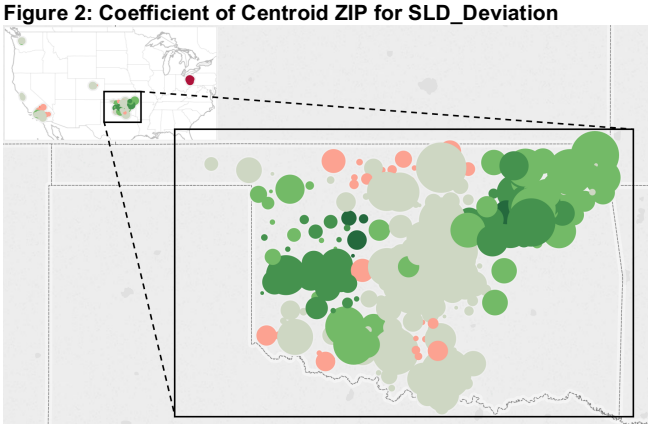
Results & Discussion

Before we jump into results and discussion, we acknowledge that we begin this study with a very challenging goal: to understand route deviations, which are essentially a human response or behavior, by only using environmental variables that describe the route. Our results confirm how challenging this is: our models are able to explain only half of the variability of the driver's behavior. However, we identify variables that show significant relationships between the characteristics of the routes, the behavior of the drivers and the performance impact of those routes.

Our models have a generalized R^2, depending on the country and the method used, that range from 0.26 to 0.35 when predicting the variable *Deviation*, and 0.37 to 0.57 when predicting the variable *SLD_Deviation*.

For instance, from the linear regression model to predict *SLD_Deviation*, we map the coefficient of the centroid ZIP variable, shown in Figure 2. Each circle represents a centroid ZIP. Size of the circle represents the number of route instances that have the same centroid ZIP, color represents the correlation with *SLD_Deviation* (red implies higher deviation impact; green implies lower deviation impact). In Oklahoma, we observe more deliveries in the northeast of the state with relatively lower *SLD_Deviation* impact. We also observe a clear cut between the center state and the rest of the state that more deviation in SLD appears due to the geographical features at the center of Oklahoma.

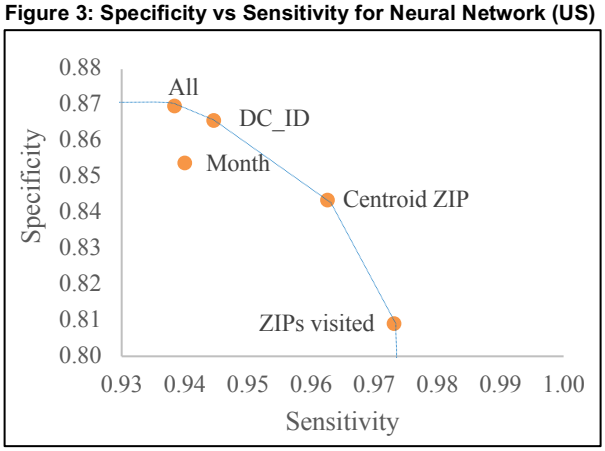
Figure 2: Coefficient of Centroid ZIP for SLD_Deviation



In general, performance of the neural network and random forest models is better than regression analysis. For the binary variable *Deviation*, the proportion of routes correctly predicted (known as accuracy) ranged from 71% to 84% depending on the model and country. To understand how important is each variable in the predictive capabilities of the model, we do a sensibility analysis. Each variable is alternatively eliminated from the model, and the performance of the model is measured. Through this analysis we find that the number ZIPs visited and the ZIP code of the centroid of the customers visited are the most significant variables.

Besides accuracy, we used sensitivity and specificity to measure the performance of the models that predict *Deviation*. Sensitivity measures the proportion of deviated routes classified as such, while specificity measures the proportion of routes classified as deviated that actually did deviate. Both variables are expected to have a negative relationship. As models over predict deviation, they increase sensitivity at the cost of decreasing specificity. In figure 3 we see an example of this happening, when performing the sensitivity analysis for the neural network model, using data from US deliveries. Starting from all the variables selected through the regression analysis, labeled as "All", we see how removing variables can increase sensitivity from 0.94 up to 0.97. However, this is done at the cost of reducing specificity from 0.87 to 0.81.

Figure 3: Specificity vs Sensitivity for Neural Network (US)



Conclusions

The data reflects an environment where delivery crews deviate significantly from the planned routes. In Mexico nearly half of the routes deviate and in the US four out of five. In this environment, following the planned sequence is unlikely to be enforced or regulated, but rather a decision that drivers take. These deviations are not adding economic value.

To explain and predict the behavior of the delivery crews, we build regression analysis and classification models that use parameters of the routes as explanatory variables. In the process of building these models we discover interesting relationships.

As routes have a larger number of customers, or visit a larger number of ZIP codes, drivers are more likely to make worse decisions: the proportion of deviated routes increase, as well as the proportional distance added to these routes. A variable created to describe the location of all the customers visited by the route, Centroid ZIP, showed good predictable capabilities. However, models that use this variable need training data of the particular regions where it will be used.