VISIBLE SUPPLY CHAIN MGMT

Isaiah Morgan, Cloe Wynn, Aaron Hill, Jonathan Linford

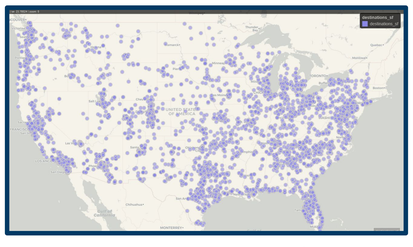
BRIGHAM YOUNG UNIVERSITY · FOUNDED BYU 1875 · PROVO, UTAH

The Problem

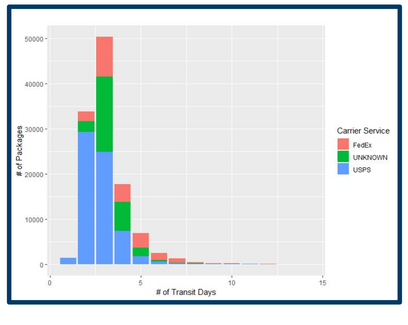
Visible Supply Chain Management provides small to medium size companies with shipping solutions to help these growing companies get their products to buyers. The main part of Visible's business is shipping and they are constantly looking for ways to provide better, faster, more reliable service to its customers. They needed a way to predict the carrier and service that would provide the fastest shipping time given the origin and destination of the package.

The Objective

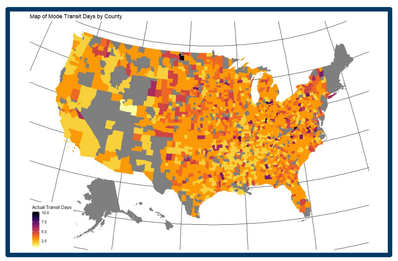
Our objective when starting this project was to work with Visible to create a model that would accurately predict shipping time of package given the carrier, service type, origin, and destination. We hoped to leverage Visible's vast amount of data that they has collected from the millions of packages they have distributed to predict the shipping time for a given destination and then verify the accuracy of our predictions.



This map depicts a small sampling of the destinations of the packages that Visible has shipped within the USA.



This plot shows the distribution of delivery days per carrier.



This graphic is a heatmap that shows the mode number of transit days to reach each of the counties in our dataset.

The Model

We developed a multiple linear regression model to predict the number of transit days for a given destination zip code. Because we cannot easily treat zip code as its own feature, we broke down zip code into two component pieces - distance and angle which are both continuous variables.

Because we are primarily building a predictive model we decided to use root mean square error as our primary indicator for success. Having the smallest error in our predictions was our top priority.

When we first built our model we had a RMSE of about 2 days. However, after we transformed our response variable and used a AIC variable selection process we simplified our model and got about a RMSE of 1.2. We further improved the model by running a k-folds cross validation technique which achieved a much lower RMSE of 0.33 that continued to decrease to .299 days as we added more rows of data. We experimented with different cross validation techniques including leave-one-out and repeated k-folds validation but both were much slower and provided an insignificant benefit over a 5-fold cross validation.

Our final model was built with 5-fold cross validation and consisted predicting the log number of transit days by looking at Origin Zip Code, the interaction of Distance Traveled and Carrier, and the sin and cos components of the angle package traveled.

The Results

Interestingly we found that the cosine component of the angle played a less significant role than the sine component. This along with the map of mode transit days seems to suggest that the East-West distance a package travels affects transit time less than North-South travel. This was an unexpected finding but gives insight into how many carrier businesses operate.

We also found that the distance a package travels is largely insignificant in predicting the number of transit days on its own. However, its interaction with which Carrier was used was a significant predictor.

One feature we examined that was not included in the final model was the type of delivery service in addition to the carrier. This feature proved to be insignificant in our predictions because Visible only used priority shipping methods in the data provided.

Unfortunately given the scarcity of our data it is difficult to give a highly accurate prediction of transit time by zip code, but we are able to give good generalizations based on broader regions. We achieved a RMSE of .299 and a MAE of .223 in our final model.

Conclusion

With more data, the model can easily become more sophisticated and accurate. The only real limitation on this project was the scarcity of useable data. However, now that the majority of the heavy lifting has been done with accessing the data, getting more data will be much easier in the future.

The model we built was based on about 115,000 useable observations and it already has a significant amount of predictive power, with 1 to 2 million rows that power would increase dramatically.

In addition, features such as weather and other delays could be added to the model allowing for more in depth analysis. Time of year could also be taken into account; as well as proximity to specific holidays.

All in all, the data collecting script we wrote, the model we built, and the R Shiny App we created provide a valuable starting place for future projects and explorations into Visibile's shipping business. Future teams will be able to utilize these resources and provide even more insightful solutions to Visible's supply chain questions.

Collecting The Data

Collecting the data for our model proved to be the most difficult part of this process. We found that we had to make major adjustments to original plan in order to account for the limitations in the API and the data provided.

Original Solution

In our original solution, we had planned to use the tracking numbers and carrier names provided by Visible to directly access the APIs of each carrier. We planned on writing a script that would loop through the tracking number provided and then call the endpoint that corresponded to the correct carrier for that tracking number.

Problems Encountered

We soon found several problems with this solution. The first was that it is difficult to receive permission to access some of the carriers APIs. While reading through the documentation of several of the carriers' APIs, we found that they were not consistent in the information that they sent back. Some carriers provides certain attributes while other carriers did not. Additionally, we were unable to gain access to several of the carriers' APIs. After running into these problems we determined that our time would be more effectively used by finding another way of collecting the data.

Final Solution

For our final solution, Visible provided us with access to an inhouse API which would allow us to collect data that consistently returned the same information about each package and did not require explicit permission to access each individual carrier's API. We built a script that we ran through an Azure VM instance which made calls to Visible's API to collect more data on each of the packages. This solution was slow but we were still about to collect information on about 350,000 packages over the course of several days.

We also discovered that Visible had provided us data from 2018. We found that many of the tracking number expired after several months so the vast majority of the data wasn't useful to us. To solve this solution, Visible was able to provide us with tracking numbers from 2019 and we were able to use that to collect our data. We also found that the API didn't support one of the carriers contained in our dataset and that other responses from the API were missing the delivery date. These issues brought the number of usable rows down to about 115,000.