UNKNOWN TO KNOWN: PREDICTING TRUCK GPS COMMODITY USING MACHINE LEARNING[1]

Mausam Duggal, Bryce Sharman, Rick Donnelly, Matthew Roorda, Sundar Damodaran, and Shan Sureshan

Introduction

Machine Learning, Big Data, and Data Mining often intersect, superimpose, and get confused with each other. The proliferation of machine learning applications and successes in the last few years has further blurred this distinction to the point where Big Data is perceived, rightfully so, to be a requirement for machine learning. While the term has been in use since the 1990s, there exists no consensus on what constitutes Big Data, especially if it were to be based on the size of the dataset alone. Gathering Big Data can be prohibitively expensive and time consuming for most private or government entities. This is especially so in transportation planning, where even a large household survey such as the Transportation Tomorrow Survey (TTS) that is conducted every 5 years and represents 5% (~175,000) of the households in the Greater Golden Horseshoe Area would not qualify as Big Data[i]. *Thus, machine learning on "smaller" datasets is a compelling application, especially in the transportation field, subject to a careful evaluation of the bias-variance tradeoff.*

To satisfy data needs for freight modelling and planning, the Ministry of Transportation Ontario (MTO) conducts the Commercial Vehicle Survey (CVS) approximately every five years. The last one, conducted in 2012, collected information on the route, commodity transported, and vehicle configuration for ~ 45,000 truck tours surveyed across 220 sites located in the Province. Since 2014, MTO has also purchased truck GPS data that show anonymized truck travel histories for trucks with at least one trip end in Ontario or passing through the Province. The GPS data is the most extensive depiction of truck movements on the Province's road network, although its market share is unknown. it Due to the anonymous nature of the GPS data collection it also fails to collect detailed information about the commodities carried or the vehicle type, which are important for freight modelling and planning applications.

WSP has built a machine learning (ML) model to explore inferring the commodity moves from truck GPS data usage on Ontario roads, as part of a project developing a freight forecasting model for the province. The model was built using the CVS and the Pitney Bowes firm data to infer the most likely commodity, aggregated to the Canadian Freight Analysis Framework(CFAF) definitions, carried by that **tour** using nearby firms, employment by industry types, and geographical locations along every stop of the tour. A separate model was also built to distinguish between an empty and commodity carrying truck **trip**, which is applied within tours. At the time of writing this paper, the empty truck trip model was yet under development and hence omitted from this publication. The remainder of this paper focuses on truck tours that carried a commodity of which the CVS contains a total of 31,491 records.

This paper is structured as follows. Section 2 describes the motivation and approach to the model. Section 3 describes the input data sources. Section 4 describes the machine learning model, its accuracy, and shortcomings. Section 5 notes future directions.

Motivation

[1] 54th Annual Meetings of the *Canadian Transportation Research Forum*, May 26 - 29, 2019 at Vancouver, British Columbia

The primary goal of the study was to cross-walk commodity flow patterns in the CVS to the GPS data. Two approaches were considered and one tested: first, a naïve sampling approach using observed CVS distributions; second, a ML model that learns the patterns in the CVS to transfer them to the GPS data.

The sampling approach relies on a Monte Carlo simulation using the spatial distribution by Standard Classification of Transported Goods 2-digit (SCTG2) observed in the CVS. It would be trivial to select a suitable geography like Census Sub Divisions (CSDs) to segment the SCTG2 distributions in the CVS. However, this would very quickly lead to the "curse of dimensionality" because a little over 329,476 (574 CSDs * 574 CSDs) statistically significant distributions would be required to cover all pairs of CSDs. The authors recognize that not all the CSD pairs are observed in the CVS; in fact, only 9,889unique CSD pairs are captured by the CVS. This would require a total of 9889 distributions, still a large number. However, the GPS truck tours are more abundant and require synthesized commodity distributions for CSD pairs that are not observed in the CVS. Other segmentations, such as employment densities are also possible. But, any attempt to increase the granularity of the segmentation is bound to lead back to the curse of dimensionality with questionable accuracy. Notwithstanding the challenges noted above, a CSD based sampling approach would also completely disregard the relationship between commodity flows and firms. Thus, this approach was not tested further. Prediction of commodity flows for many commodity classes is a problem well suited to the field of supervised learning, specifically classification[ii]. The goal is to construct a succinct ML model using the CVS that can predict the commodity carried in the truck as a function of attribute variables such as geography, tour distance, nearby firms to start and end locations, etc. of the individual truck tour. Once constructed, the ML model would be applied on the GPS data to predict a commodity for the entire truck tour. The ML model was tested given the ability to build spatial and firm type relationships.

Input Data

Commercial Vehicle Survey

The 2012 CVS, like its predecessors, intercepts trucks at data collection sites across the Province of Ontario highway network. The CVS, on average, achieves a 4% sample of commercial vehicles passing each data collection site, which represents 10% effective sampling rate after considering routing of the sampled vehicles, although significant variation is seen across data collection sites. Each truck tour (consisting of the commodity origin, intermediate stops, commodity destination, truck configuration, tour length, etc.) is tagged with a single SCTG2 commodity classification. A total of 42 SCTG2 codes can be found across the 31,491 truck tours, excluding an empty tour. The CVS, however, does not collect any information on the shipping and receiving firms. Furthermore, many other variables that it collects such as truck configurations, truck type etc. that are very useful for freight planning applications cannot be used in the ML model because of the lack of similar information in the truck GPS data. The initial set of data from the CVS that could be used are shown in Table 1.

Table 1. Initial set of variables from the CVS for the ML model

Variable	Description
Commodity class	1-43: SCTG2 1-11: CFAF Groups
Number of stops in tour	Limited number of stop information is collected by the survey, including trip origin, commodity origin, previous stop, border crossing, counting station, next stop, commodity destination, trip destination
Lat/long of each stop	Tagged to CSD
Weekend vs Weekday flag	Tagged as 1 if survey was conducted over weekday; 0 otherwise
Tour length	5km to 7500 km

Pitney Bowes Firm Data

The 2012 Pitney Bowes firm data was obtained by the MTO during the development of the Province of Ontario's passenger and freight forecasting / travel demand model (TRESO). This data set includes information on all the firms (390,162) in the Province, including its NAICS, size, and XY coordinates. Table 2 shows a sample of the database.

Table 2. Sample Format of Pitney Bowes Firms Data

Firm_ID	CSDUID	NAICS	SIZE (emp)	Latitude	Longitude	EMPLOYEES
7903	3501012	23	1 to 5	86.15823	16.624465	3
7904	3501012	23	20 to 50	86.15828	16.624467	35

Socioeconomic Data

The MTO has developed a detailed estimate of GDP (gross domestic product) by NAICS2 for each CSD in the province. The data spans from 2011 to 2071 in one-year increments and includes population estimates by age and sex.

The Learner

Our approach to data preparation and data mining - two primary stages of a learner are noted below.

Data Preparation (Cleaning, Integration, Selection, Transformation)

Truck flows and commodity flows follow a rational pattern that reflects the underlying socioeconomic mix of firms, people, and prevailing economic conditions. Thus, the origin and destination of each CVS tour can be assigned to a spatial unit, and the attributes of those spatial units can be used to predict the commodity moved on the tour. The choice of spatial unit is influenced by the following criteria. First, it should be possible to generate credible aggregate level attributes for this spatial unit for training the model. Second, the spatial unit should allow for a meaningful evaluation of the model's accuracy while proving useful in policy planning. The CSD geography (574 CSDs in Ontario) satisfied the above criteria.

Classifying 42 SCTG2 classes is a challenging task for any learner, especially given the small size of the CVS dataset. Thus, a more tractable specification was desired. This was achieved by re-mapping the SCTG2 to the CFAF commodity groups, for a total of 11 CFAF groups. Figure 1 shows the distribution of CVS survey records mapped to the CFAF groups.

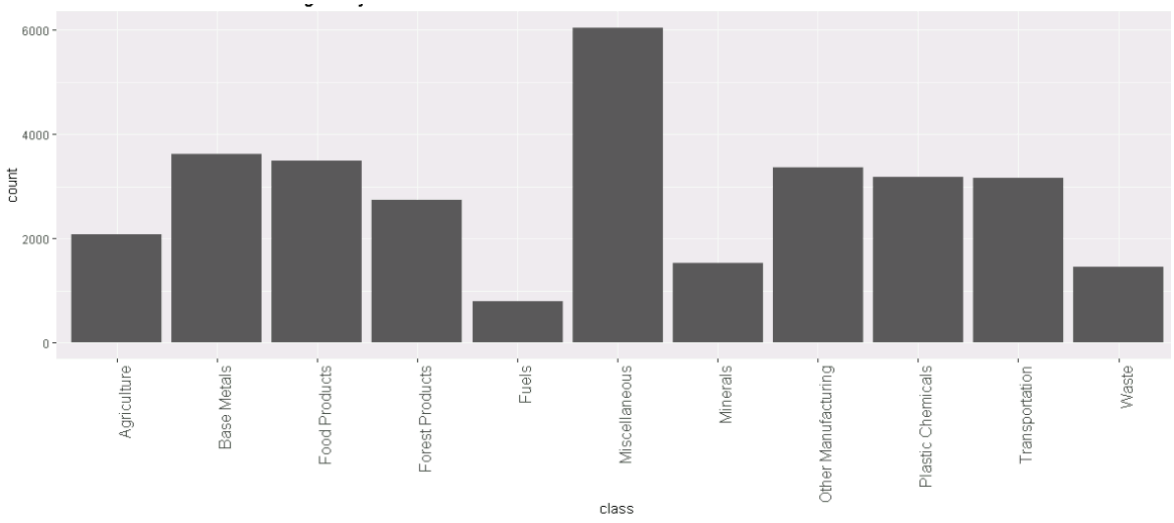

Figure 1. Number of CVS Surveys by CFAF Groups

The CVS survey did not collect any information on the firm that the tour originated from or is destined to. However, this is a significant source of information for the learner. Any simple map matching process that attempts to exactly tag every CVS tour with a Pitney Bowes firm would achieve limited success because of differences in spatial geo coding accuracy. For example, a truck driver reports either an exact address, postal code, intersection, or landmark. This information is converted to latitude/longitude format in real time using a mapping software integrated within the survey instrument. In all cases, the resulting tour information is map matched to the MTO's road network. The firm data on the other hand is either tagged to the road network or building polygon depending upon whether it is an urban, suburban, or rural setting. Furthermore, even an exact address obtained in the survey could not be mapped to the Pitney Bowes firm dataset with a high degree of certainty. Incorporating firm information in the estimation data was critical, hence a simulated annealing approach was developed that would help identify a **set of firms**(probabilistic) that a truck could have serviced at each stop on the tour. Figure 2 presents a schematic of the simulated annealing (SA) approach and Figure 3 shows the Gaussian distance and probability calculations in the SA.

Initialize - Generate buffer around first CVS stop:
250m: Greater Golden Horseshoe (GGH) and Ottawa region;
500m: all others;
Starting variance: user defined (5^{th} root of buffer: default);

While (no Pitney Bowes firm found in buffer):
Increase buffer size by 250m radius;
Compute Gaussian Distance from CVS point to each Pitney Bowes firm segmented by NAICS;
Compute weighted probability for every NAICS category in buffer;
Collect number of employees by each NAICS in buffer;

If (buffer radius reaches 4000m with no Pitney Bowes firm):
Exit
Return to next stop in tour

Figure 2. Framework for the Simulated Annealing Algorithm

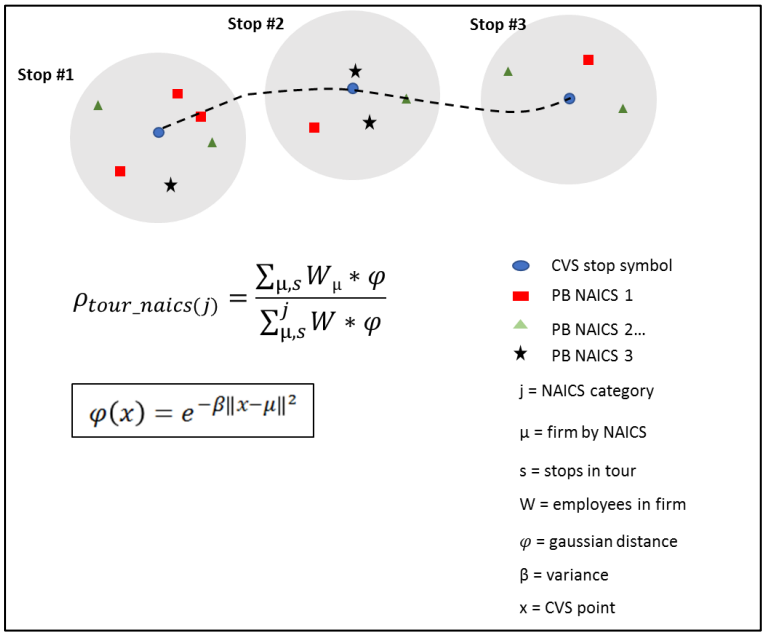

Figure 3. Gaussian Distance and Weighted Probability by NAICS – Simulated Annealing

Table 3 shows the final set of attributes used for training the ML model.

Table 3. Final Attribute Set in ML model

Attribute	Cols	Source	Description
Total tours in estimation data			31,491
CFAF group	1	CVS	11 groups
Tour length	1	CVS	Kms
Ln(tour length)	1	CVS	Natural log of tour length
Weekday flag	1	CVS	1 for weekday; 0 otherwise
First_csd	1	CVS	CSD of the first leg of the tour
Last_csd	1	CVS	CSD of the last leg of the tour
Prob(naics)	22	SA + PB	Weighted probability for each NAICS across the entire tour found in tour stop buffers
Emp(naics)	22	SA + PB	Number of employees for each NAICS across the entire tour found in tour stop buffers
First_gdp	1	SA + SED	GDP of the CSD that belongs to the first leg of the tour
Last_gdp	1	SA + SED	GDP of the CSD that belongs to the last leg of the tour
First_pop	1	SA + SED	Population of the CSD that belongs to the first leg of the tour
Last_pop	1	SA + SED	Population of the CSD that belongs to the last leg of the tour

CVS: commercial vehicle survey; SA: simulated annealing; SED: socioeconomic data; PB: Pitney Bowes

Data Mining - Gradient Boosting Machine

Two learners, Deep Neural Networks (DNN) and Gradient Boosting Machines (GBM),were evaluated. Given the number of categorical attributes (First CSD, weekday flag etc.), the GBMs were carried forward after the initial round of testing. Characteristics of the final GBM are shown below in Table 4.As per convention, the total commodity truck tours in the CVS were split into training and validation data using a stratified sampling technique on the CFAF groups. Given the relatively small dataset being used in the development of the ML, *k-fold cross validation*[iii] was used to overcome problems associated with over fitting and bias. The hyper parameter tuning column in Table 4 identifies the parameters that were optimized using a random search pattern.

Table 4. Final Characteristics of the GBM

Characteristics	Value	Hyper parameter tuning
Training data (no empty truck tours)	25,188 tours	
Validation data	6,303 tours	
Number of commodity classes (CFAF groups)	11	
Number of folds (k-fold)	5	
Number of trees	250	Yes
Row sampling rate	0.67	Yes
Column sampling rate	0.67	Yes
Max depth of tree	100	Yes
Number of bins (continuous data)	900	Yes
Number of categorical bins	900	Yes
Weights*	1 (no weighting)	

*The training data was not weighted to account for any imbalances at the time of writing this publication.

Gradient Boosting Machine Performance

The GBM's performance was evaluated on the validation data. Two key metrics were chosen to evaluate the GBM with the following in mind: *Which CFAF groups are alike, causing the GBM to cross-classify tours between them? Is the GBM better than simply guessing randomly based on frequency of each class?*

Two metrics were chosen that would provide insight to the above questions: first is a confusion matrix (CM)[iv] while the second is the Cohen Kappa (CK) metric[v]. In a CM, each row of the matrix represents instances in an actual class (observed) and each column represents the instances in a predicted class. In a perfect classifier, all non-diagonal entries will be zero. The CK is more robust than just a simple percent agreement (sum of diagonals in the CM divided by the total of the CM) calculation obtained from a CM, as it considers the possibility of agreement occurring by chance. There is some ambiguity in interpreting the values of CK, but in general anything less than 0.2 is questionable and close to 1.0 is a perfect model. Table5 shows the CM obtained from the validation data. Some observations are presented below:

- The overall accuracy of the model is at ~39% (1-0.6110)[vi]
- The MNRLS (minerals) category, has the highest precision and recall rates[vii], followed by MISC.
- MISC (mixed freight) is often confused with the other CFAF groups. This is partially because MISC is the majority class in the training dataset. Second, the SA algorithm is unable to generate attribute values for the MISC category that are distinct from other CFAF groups given the ubiquitous nature of MISC.

Table 5. Confusion Matrix on validation data – GBM

▾ PREDICTION - CONFUSION MATRIX ROW LABELS: ACTUAL CLASS; COLUMN LABELS: PREDICTED CLASS

	AGRI	BMETL	FOOD	FRPAP	FUELS	MISC	MNRLS	OTHMF	PLCHM	TRANS	WASTE	Error	Rate	Recall
AGRI	130	51	54	16	1	82	13	29	26	9	7	0.6890	288 / 418	0.47
BMETL	17	288	59	17	7	123	22	77	47	47	20	0.6022	436 / 724	0.31
FOOD	23	64	280	28	0	162	5	57	52	22	6	0.5994	419 / 699	0.34
FRPAP	6	65	71	156	2	106	12	47	52	23	8	0.7153	392 / 548	0.46
FUELS	4	20	7	3	60	22	9	11	13	7	4	0.6250	100 / 160	0.66
MISC	28	114	122	32	8	668	9	95	61	65	9	0.4484	543 / 1,211	0.40
MNRLS	8	37	18	10	2	26	171	10	9	6	11	0.4448	137 / 308	0.63
OTHMF	21	96	71	23	4	164	10	163	61	55	6	0.7582	511 / 674	0.25
PLCHM	20	100	77	30	3	139	6	79	149	27	6	0.7657	487 / 636	0.29
TRANS	9	61	40	12	1	118	5	69	24	290	4	0.5419	343 / 633	0.52
WASTE	8	43	24	11	3	53	8	17	20	8	97	0.6678	195 / 292	0.54
Total	274	939	823	338	91	1663	270	654	514	559	178	0.6110	3,851 / 6,303	
Precision	0.31	0.40	0.40	0.28	0.38	0.55	0.56	0.24	0.23	0.46	0.33			

The CK value and expected accuracies from the validation data are presented in Table 6.

Table 6. Cohen Kappa Metric Performance - GBM

Metric	Minimum	Value	Evaluation
Cohen Kappa (CK)	0.2	0.3	Pass
Accuracy by random chance	NA	12.2%	GBM is three times more accurate

Several plots and statistics have been computed to measure the accuracy of the GBM at a more spatially disaggregate level, keeping in mind the planning and policy applications. For reasons of brevity only two plots are presented in Figure 4 and 5 that showcase the GBM's performance. Figure 4 aggregates the results to the tour's CD (Census Division) of origin instead of CSD for ease of comprehension. Figure 5

presents the GBM's performance segmented by 100 km distance bins. The y-axis in both plots notes the percent accuracy and values over each bar are the number of observations in each category.

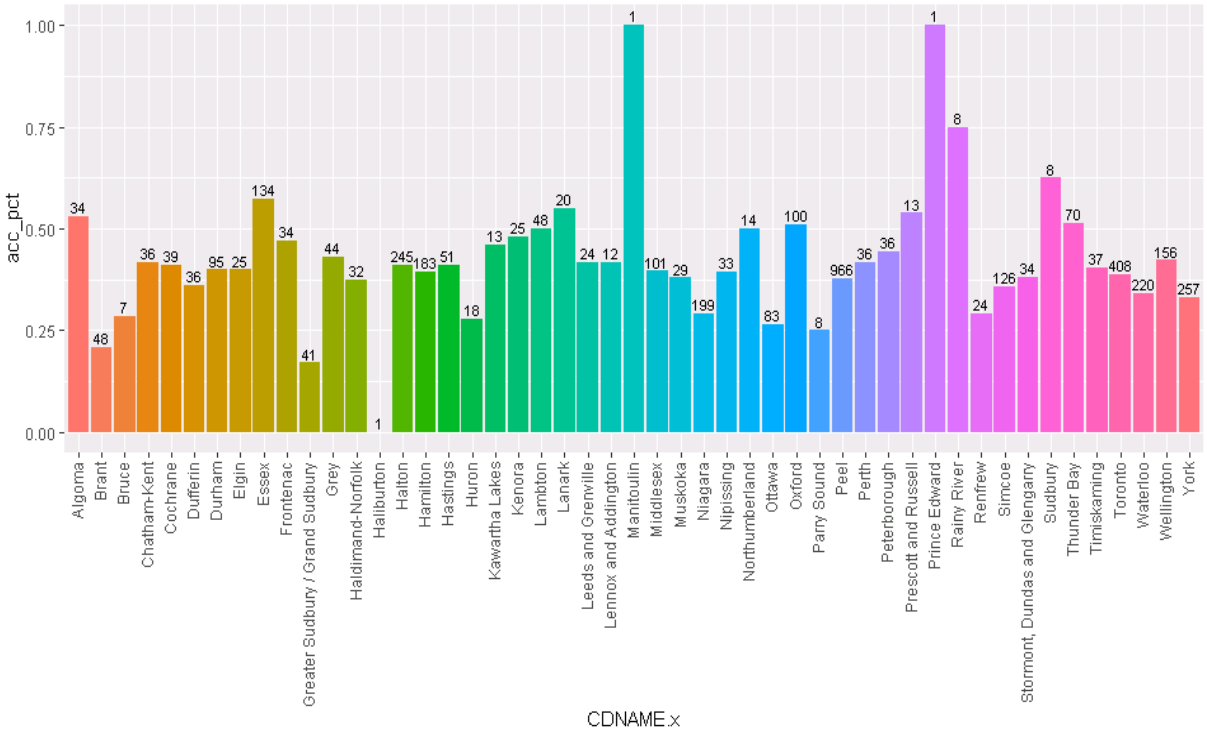

Figure 4. GBM Performance on validation data by CD of Truck Tour Origin

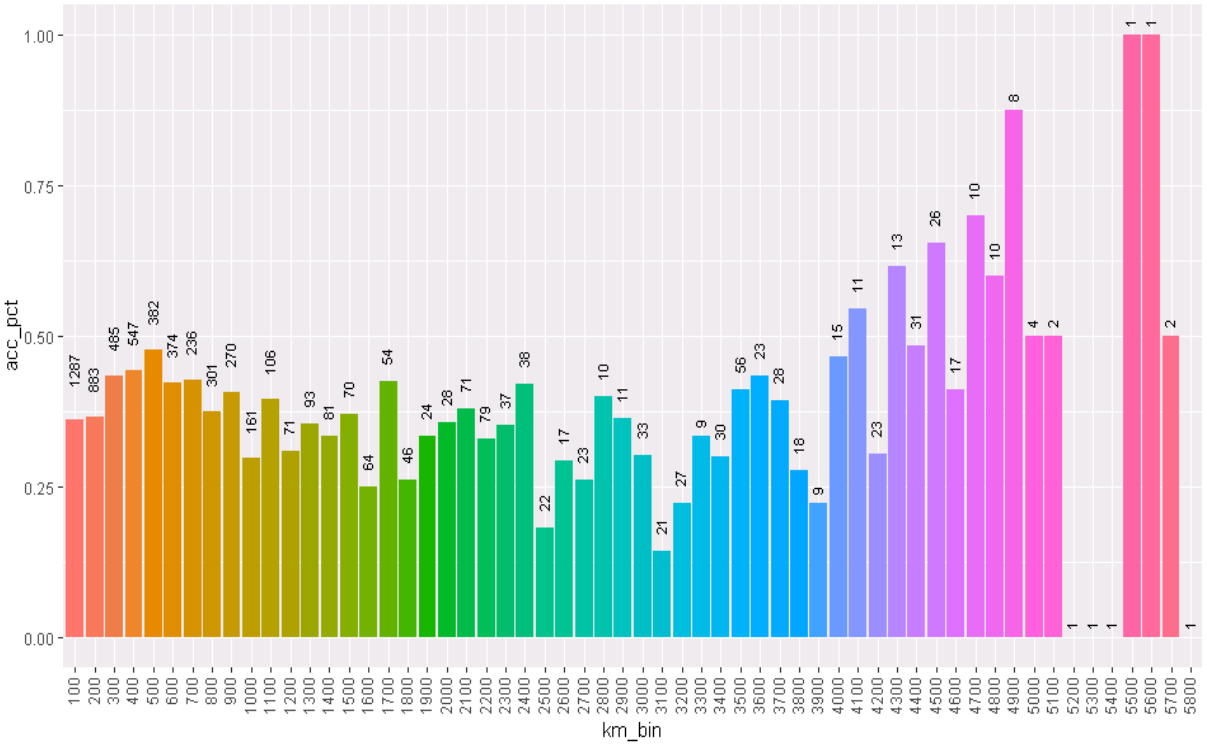

Figure 5. GBM Performance on validation data by Distance Bins (100 km)

Reflections and Future Directions

- *Can patterns learned from small data like the CVS be readily applied to the GPS?*

The MTO is collecting the next round of CVS data. Assuming a similar effort such as the 2012 CVS, there is the potential to double the training and testing data. An evaluation of the GBM model's performance (Figures 4,5) would also help focus survey efforts by geography, distance, and commodity stratifications. Thus, while the dataset used in this round of the GBM was relatively small, it is expected to help optimize future surveys and prove to be a valuable source of information in policy planning.

- *Are there any biases in the commodity flows observed in the CVS given the predominantly long-distance truck tours in the survey?*

About 15% of the CVS tours are less than 45 kms long; a distance definition borrowed from long distance passenger models given the lack of one to define long distance truck travel. This was contrary to the general expectation going into the project that the CVS predominantly captures long distance truck travel. This mitigates some, if not all biases the learner built on the CVS may have when applied to the GPS that does not favor long or short distance truck tours.

- *Do trucks with a GPS on-board behave any differently than those that don't?*

An evaluation of CVS trucks that carried a GPS unit versus those that did not showed a nearly identical distribution of commodities. This supports our expectation that the GPS carrying trucks do not behave differently than those that did not and the learner could be safely applied on the GPS truck tour data.

This learner (GBM) and its supporting package of data preparation tools are expected to be deployed in a repetitive learning environment. New CVS surveys, firm data, and socioeconomic information can continually feed the learner improving its accuracy and expanding its capabilities. Unlike standard econometric techniques this is a self-learning mechanism that although originated in the world of "Small Data" has the potential to evolve its way to "Big Data".

---

[i] Image Net, which is the most extensive Big Data source for Machine Learning has around 14 million images currently

[ii] Machine learning algorithms can be broadly classified in to two classes: regression and classification. Classification is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y). The output variables are often called classes or labels.

[iii] https://en.wikipedia.org/wiki/Cross-validation_(statistics)

[iv] Confusion Matrix: https://en.wikipedia.org/wiki/Confusion_matrix

[v] Cohen Kappa: https://en.wikipedia.org/wiki/Cohen%27s_kappa

[vi] Error rate in the CM is calculated by dividing the sum of all non-diagonal cells by the total records in the testing data

[vii] Precision is obtained by dividing the diagonal with the actual observations e.g. MISC = 0.55, which is obtained by 668/1211. Recall is obtained by dividing the diagonal with the predicted observations e.g. MISC = 0.4, which is obtained by 668/1663. Precision and Recall are bounded between 0 and 1 with a higher value reflecting a perfect classifier.