**Data Description**

The two data-sources used for this study include:

1. **The Enplaning Passenger Survey responses:** This data-source guides our understanding of the mode-choice used by the enplaning passengers. The survey responses were available from 2014 until 2019. The spreadsheet with data consisted of over 10000 responses with 65 columns. The cleaning process consisted of removing rows with null values, like for field ‘Airline’ which denotes the airline on which the passenger was traveling, mode choice to airport, etc. Responses were removed where the current flight was a connecting flight to capture enplaning passengers getting on from SeaTac airport. Further, the data was transformed and new features were created to aid in the mode choice estimation process. A boolean-type column ‘Domestic’ was created to reflect, whether the flight was domestic or international based on the value from the ‘Airline’ column. A numerical column was created to capture the luggage with a passenger; this includes the carry-on as well as check luggage. Another Boolean type feature captured whether the trip was a business trip or not. Age was calculated from the response to the question “when were you born?” and erratic values removed. The feature was further scaled to aid with the estimation, such that the maximum age value was 1 and other ages were numbers between 0 and 1. A categorical feature, representing income group was created by re-grouping the household income response to 5 categories. Month of travel was extracted from the survey date to be used as another feature. The responses were then divided into separate datasets for each year, and we ended up with around 700 responses for each year from 2014 to 2019, with 6 predictors, namely - ‘Domestic’, ‘luggage’, ‘business’, ‘Age’, ‘income’ and ‘Date\_month’. Finally, the dependent variable, mode-choice was inferred based on the combination of answers to few questions and combined to make the prediction reasonable. Based on discussions with the SeaTac staff, the following modes are chosen for estimation:
   * **Transit+:** This represents all the high-volume modes, including public transit like buses and light rail as well as airport shuttles and charter buses.
   * **Parked:** This category reflects the private vehicles parked at the SeaTac airport. This category also includes free-float car sharing operator vehicle from ReachNow and car2go.
   * **Curbside:** This category represents the curbside drop-offs by private vehicles, possibly family-members and friends dropping the passengers off to the airport.
   * **TNCs+:** This category captures the TNCs, taxis etc. drop-offs at the curbside.

Figure 1shows the variation in mode-share at SeaTac for enplaning years from 2014-2019 for the aggregated modes as described above. The data was converted to percentages to offset the effect of fewer observations so far for year 2019. We see a an increase in the use of TNCs+ mode over the years. The 2019 numbers probably do not tell the whole story as the summer months are ahead of us, when people are more likely to choose transit than in winter months. Regardless, we see a consistent competition for the limited curbside drop-offs over the years, and the problem will only exacerbate over time as the enplaning passenger volume out of SeaTac increases. This data is used for mode-choice logit model, which is discussed later. One of the key challenges, in trying to classify data like this is the class-imbalance, which is quite noticeable in our case, since the two categories Transit+ and Curbide dominate the outcome.

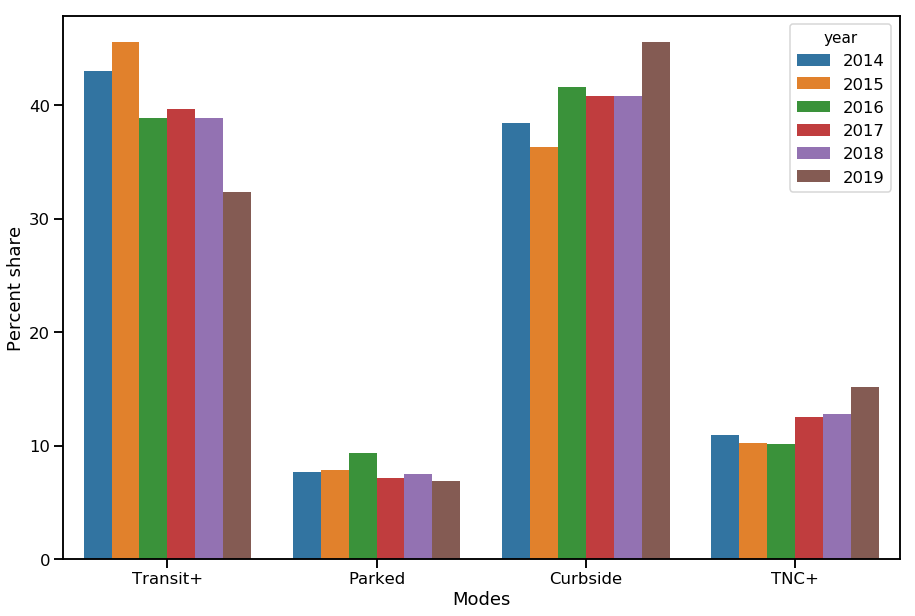
****

Figure 1: Mode-share or enplaning passengers at SeaTac from 2014-2019.

1. **Yearly Passenger Counts:** This data source is a spreadsheet with passenger count data from 1990 to 2018. The enplaning passenger count is used to forecast the future enplaning passenger count. The passenger forecast model and results are discussed later.

**Model Estimation Methods**

1. **Mode Choice Model:** The aim of the mode-choice model is correctly predict the mode-choice given the passenger demographic information etc. As described before, the enplaning passenger survey data was cleaned and transformed to give 6 features was mode-choice modeling. The modes are encoded as numbers from 0 to 3 representing each mode and the modeling task therefore is similar to classification. Several classification algorithms were used, namely, multinomial logistic regression, decision trees, random forest, gradient boosting machine, support vector machine, and Keras deep learning with Tensorflow backend.

Figure 2shows the variation of accuracy of prediction for training data from 2014 to 2019 using several algorithms. We see the multinomial logistic regression does not do well with accuracy below 50% for most cases. The poor accuracy can be mainly attributed to the class-imbalance. We see that decision tree does best with accuracy over 90% in most cases. Keras deep learning library with Tensorflow backend using a sequential model with 4 layers with 250 neurons each and 2000 epochs does a close enough job, with random forest following closely.

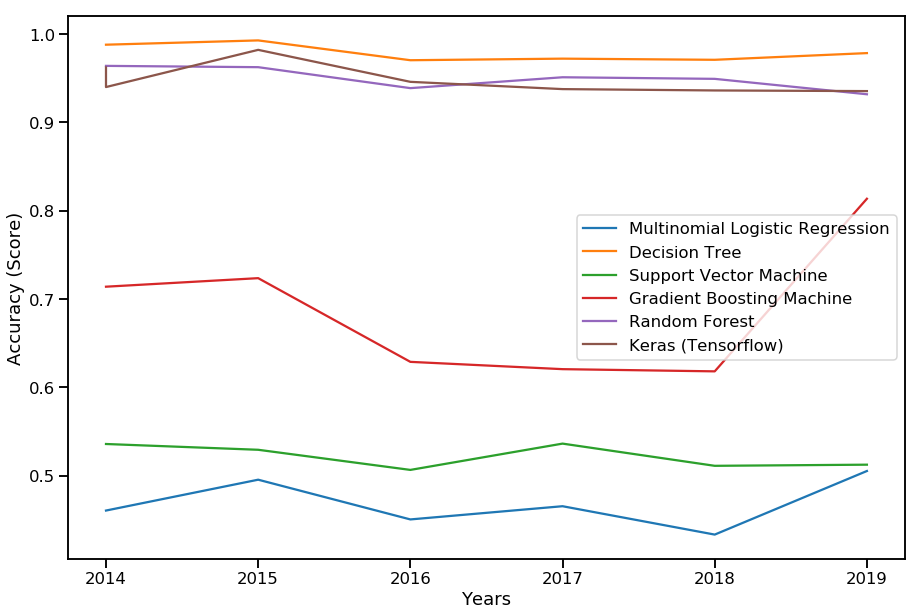


Figure 2: Comparison of modeling accuracy using training data for mode-choice modeling using various classification algorithms