Project Report

Modeling Air Travel Demand between Two Cities

Aparna Agrawal, S M Ferdous, Apoorv Maheshwari, Luis Zertuche
agrawa42@purdue.edu, sferdou@purdue.edu, amaheshw@purdue.edu, lzerttuc@purdue.edu

December 12, 2015

1 Introduction

With the increasing population and diminishing resources, it is necessary for the policy makers to efficiently plan for the future investment of the resources. One such area is transportation where policies are needed to be developed keeping next 10-20 years in mind. Much of the transportation policy decisions are taken based on the travel demand. Travel demand between any two cities is defined as the number of people traveling (or expected to travel) between the cities. The travel demand is dependent upon number of socio-technical factors, for example, population of the cities, distance between the two cities, economic condition of the cities, etc. In this project, we will mainly focus upon the demand in the air transportation sector. Our goal is ***to model the air travel demand between any two cities, based on the socio-technical factors, using machine learning techniques.*** To reduce the complexity of the problem, we will restrict our analysis to top 30 airports (enplanement-wise) of the US domestic air transportation network.

This demand model will not only give useful information on which airports (or cities) to focus while making policies for tomorrow but will also give insight on the evolution of the US domestic air transportation system. We will only use *publicly available data* in this analysis so that the policy makers who might not have access to the proprietary information of various service providers (such as airlines), can also make use of this model for future planning.

In Chapter 2, we provide background on the research work related to the area. In Chapter 3, we discuss the process of preparing the machine learning model from data collection to the model selection in detail. Then, we discuss three experiments that were performed on the generated model in Chapter 4. We describe the model performance in Chapter 5 and conclude with the main findings and future directions in Chapter 6.

2 Related Work

Since the problem of estimating air travel demand is intimately related to policy and profitability, there have been considerable quantitative modeling efforts in this domain. One of the

earliest attempts in the literature is made by Taneja [1971], which employs measures of socio-economic characteristics of airline passengers and transport related features to model and forecast total transatlantic air travel demand. The model developed by Taneja [1971] is only concerned with air traffic volume between the US and Europe, therefore the scope and treatment of the data differ from the objective in this project; however his analytic foundation appears as a good start.
A more recent attempt using similar regression tools is made by Bafail et al. [2000]; this reference provides another insight into the selection of process of explanatory variables, and the potential problem that arises with when there is multi-colinearity between predictors. Another insight from Bafail et al. [2000] relates to the economic features of the traveler that can be selected as co-variates; it argues that air travel is highly elastic to income changes of the passengers, in other words changes in income positively correlates with changes in demand. Conversely, price of the air fare behaves in-elastically, or not correlated, with demand. Caution should be taken as this reference deals with international air travel to Saudi Arabia which may not allow applying such assumptions to the current project regarding domestic US flights.
In a more advanced treatment of the similar prediction problem Kotegawa [2012] incorporates the information contained in the air transportation network structure to the models. The aim of such work is to predict how changes in the network can impact the capacity of the air transportation system.

3 Work Flow

- **Data Collection, Cleaning & Feature Selection**: The first step of the analysis was to decide the features and then collect data for each of them. As one of the conditions was to only use the publicly available data, that also constrained our feature selection. Following is a list of all the selected features along with the respective data sources.

  - **Demand Data:** Market demand data for all the routes between the 30 airports has been collected from the BTS Database for the years 2005-2014. The raw data has been processed and compiled in the required format for the analysis.
  - **Distance:** The distance information between any two airports has been calculated from both the BTS data and the latitude-longitude data; for the computation, the *geosphere* R package was used, and distances were verified with Google maps.
  - **Population:** The population data for each city where the airports are located has been collected from pop.
  - **Economy Metric:** Per capita income of the cities is collected from Department of Numbers cap and National GDP is collected from Wikipedia for the years 2005-2014.
  - **Type of City:** A city has been categorized as in two categories: Industrial and Tourist. We decided on these based on the group discussion and some consultation to the city revenue data available online.
  - **Type of Airport:** We compare the list of 30 large hubs in the FAA data-set with the hubs listed by the three largest full-service airlines in US, viz. United Airlines (UA), American Airlines (AA) and Delta (DL). Based on this comparison, we finally select

a set of 21 airports that are fixed as hubs in this work. Additionally, we add Chicago Midway to this list because of its large volume of operations, resulting in a final list of 22 hub airports.

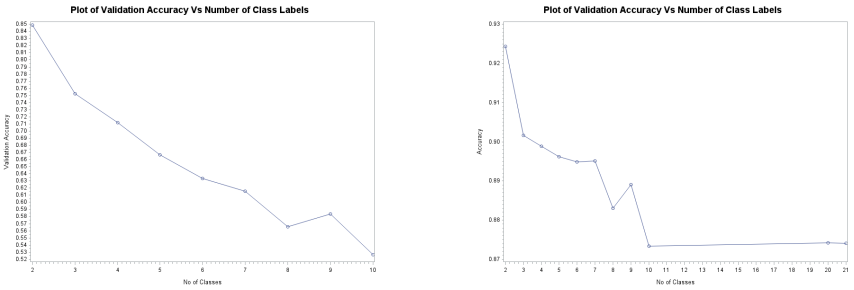
The above data sets were manipulated using the *dataframe* type in R. The final data frame contains the above listed variables as columns and each row identifies a route.

- **Preliminary Analysis**: It is always a good practice to study the data distribution of various variables employed in a model. As a preliminary step, we made scatter plots of the predictor variable (Demand) versus all the explanatory variables. Also, before applying any other more sophisticated technique, multiple linear regression studies were conducted on the data set. Regardless of input variable transformations, the best linear model fit that was achieved corresponded to an adjusted **R-squared of 0.2**. Hence it became obvious that the multiple linear regression can no longer be performed to predict the demand values. To remedy this, we converted our problem to a classification problem and divided the demand data in different categories.

- **Labeling the data**: Adding to the results of the regression, we also understood that the point estimate of the demand practically makes no sense. The more meaningful quantity is the range in which the demand might lie. Thus, we labeled our demand data into multiple categories by dividing an ordered set of the demand into the given number of categories by putting equal number of data points in each category. A contrasting approach could be to divide the data in categories with same range but we did not use this approach to make sure that we have enough training data points for each category. The latter approach will be more suitable for a real world complex problem with a large amount of data.

- **Algorithm selection** : Motivating from the fact that predicting actual demand value is not worthy, we switched to classification. As learning algorithm we experimented with two discriminative model such as Neural Network and Support Vector Machine (SVM) as well as generative model such as Naive Bayes. After analyzing the results, we fixed our learning algorithm as Support Vector Machine. The Naive Bayes which has linear decision boundary performed poorly on the other hand the performance of Neural network in our setting was very architecture dependent.

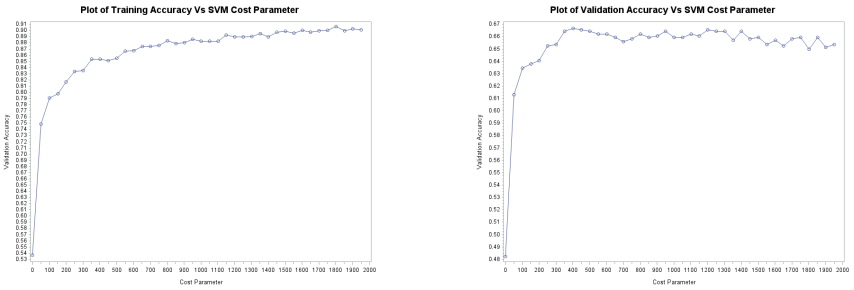
- **Number of categories** : To identify optimal number of categories for our data, we ran SVM with Radial Basis Function (RBF) kernel on training as well as our validation data set. After obtaining the validation accuracy, we decided that the number of classes should be fixed to **five**. The variation of validation accuracy with number of categories is shown in the Figure 1. Note that we have used two methods to calculate the accuracy: a) Exact Match, and b) Adapted Hamming Loss. The exact match indicates the percentage of samples that have all their labels classified correctly. On the other hand, the latter heuristic calculates the distance between the gold label and the predicted label and normalizes it with the number of categories. The exact match is the strictest metric whereas the adapted hamming loss is a relaxed metric. We can observe in the plots that the validation accuracy keeps on decreasing for the exact match metric with the increase in the number of classes (53% for 10 classes) whereas the validation accuracy converges to $\sim 87.5\%$.



(a) Exact Match (b) Adapted Hamming Loss

Figure 1: Variation of the Validation Accuracy with the number of categories

- **Tuning parameters** : The only parameter that needs to be tuned in our model is the cost variable. We ran our training model with different cost values and the cost value of **1000** gave highest training and validation accuracy. The plots are shown in the Figure 2.



(a) Training Accuracy (b) Validation Accuracy

Figure 2: Tuning of *Cost* hyper-parameter for 5 classes

4 Experiments

4.1 Transition Prediction

One of the motivation behind this study was to predict the air demand dynamics. The way of understanding how demand changes is to observe the transition between demand classes. By transition we mean the point when demand changes from one class label to another. We experimented transition prediction with our model. We increase the number of classes from 2 to 10 and for each class we only keep those data as test set which have different label than the previous year. The hyper parameter C was set to 1000. The Fig. 3 shows the plot of number of class vs. Accuracy of the prediction. The prediction accuracy falls down with the number of classes as expected. The prediction accuracy is quite good if we limit our number of classes as 5 or less. The accuracy of the prediction is around 68% for the 5 classes which is as representative as the usual test set.

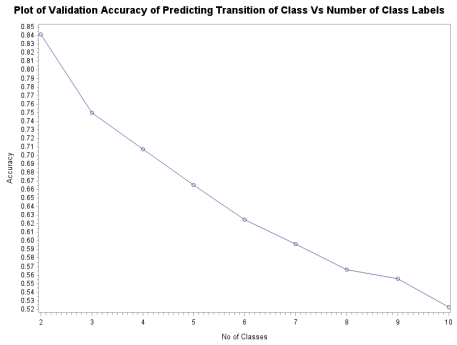


Figure 3: Transition Prediction Accuracy with the number of categories

4.2 Ablation Study

We performed ablation study on our data to identify the most important features which we should retain in our model. To do so, we ran SVM algorithm on our data by removing one parameter at a time and it was observed that the best results on validation data set were obtained if US economy is removed from the model. The results of training accuracy and validation accuracy which we obtained for five categories of demand data are summarized in the Table 1.

Variable omitted	**Training Accuracy**	**Validation Accuracy**
Distance	0.819	0.571
Origin Income	0.830	0.600
Destination Income	0.835	0.631
Origin Population	0.824	0.581
Destination Population	0.842	0.632
Hub Flag	0.858	0.630
US Economy	0.827	0.704
Tourist Flag	0.857	0.614
Industry Flag	0.871	0.617

Table 1: Results from the Ablation Study

4.3 Variation of Demand with Distance

Another interesting experiment was to see how demand changes with the change in distance. Our hypothesis was that the demand will be higher for the routes with large distance and lesser for the routes with small distance. This can be attributed to the fact that for small distances, other modes of transportation dominate over air travel. But at a particular distance, there will a sudden shift in the preference towards the air travel and thus, we should see a non-gradual change in the air travel demand at this tipping point. For this study, we chose the data point with the maximum demand in our data set (Los Angeles to San Francisco) and varied the value of distance from 0 to 10,500 km in it, keeping all the other variables same. The results from the experiment are shown in the Figure 4. In the figure, the black lines represent the range of the predicted category. We can see that demand is low (category 2) initially but there is a sudden jump (to category 5) at 500 km ($\sim 300mi$). Please note that the category 5 has a large range due to our categorization method. This can be avoided with the help of sufficiently large data set and division in categories with equal

range. We also observe an unexpected decline in the demand for very large distances. We believe that this happens because our model was trained only using the data from the US domestic market and these large distances doesn't exist in this market, Thus, the results for very large distances should be regarded as inconsequential.

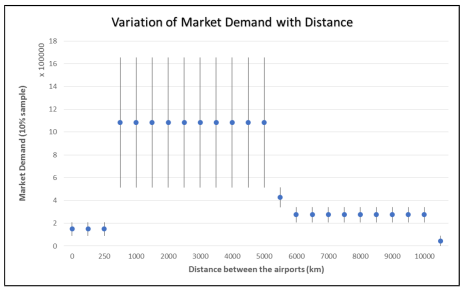


Figure 4: Results from the Distance Experiment

5 Model Performance

The final feature set selected after the ablation study are **distance, population of two cities, type of city as Hub, tourist and industry, per capita income of the two cities**. We omitted US Economy feature from our final feature set. The learning algorithm is SVM with radial basis kernel function. There is only one hyper parameter of our model which is the Penalty or Cost. We set it as 1000. With this setting, we run SVM with different permutation of the data. We generate 10 random permutation of the whole data. For each ordering the first 80% data are chosen as training set and the rest are chosen as test data. We recorded the model performance on each permutation. The average accuracy on training data is **83.5%** with standard deviation of 0.27%. On the other hand the test accuracy is **71.1%** with standard deviation of 1.2%. The main interpretation of the result is that the developed model is stable as for over 10 runs on random permutation on data, we observe small variance on performance.

6 Conclusion & Future Directions

The group project has been a great learning experience for the team and in the end, we were able to prepare a stable and sufficiently accurate model of the air travel demand between two cities using SVM. In the process, we also identified some future directions for this work as listed below:

- **Equal Range Categories**: The machine learning model should be developed with categorization with equal range of categories. This approach will be more practical for real world large data set.
- **More airports**: To restrict the scope of the problem, we only focused on the top 30 US airports but this model can be extended to more airports.
- **Application to ongoing research work**: This demand model will be adapted to extend the discrete choice model given by Sha et al. [2015].

References

RITA-BTS-Transtats. URL http://www.transtats.bts.gov/databases.asp?mode_id=1&mode_desc=aviation&subject_id2=0.

Per Capita Income Data . URL http://www.deptofnumbers.com/income/metros/.

OpenData, Airport Codes mapped to Latitude and Longitude in the United States. URL https://opendata.socrata.com/dataset/Airport-Codes-mapped-to-Latitude-Longitude-in-the-/rxrh-4cxm.

Population data . URL http://factfinder.census.gov/.

Abdullah O Bafail, Seraj Y Abed, SM Jasimuddin, and SA Jeddah. The determinants of domestic air travel demand in the kingdom of saudi arabia. *Journal of Air Transportation World Wide*, 5 (2):72–86, 2000.

Tatsuya Kotegawa. Analyzing the evolutionary mechanisms of the air transportation system-of-systems using network theory and machine learning algorithms. 2012.

Zhenghui Sha, Kushal Moolchandani, Apoorv Maheshwari, Joseph Thekinen, Jitesh H Panchal, and Daniel A DeLaurentis. Modeling airline decisions on route planning using discrete choice models. In *15th AIAA Aviation Technology, Integration, and Operations Conference*, page 2438, 2015.

Nawal K Taneja. A model for forecasting future air travel demand on the north atlantic. Technical report, Cambridge, Mass. Massachusetts Institute of Technology, Flight Transportation Laboratory,[1971], 1971.