Quantifying Impact of Weather Condition on Travel Time

Submitted By-

Sambhavi Joshi
MS-GIST, 2020
Georgia Institute of Technology

Abstract

Most transportation systems operate at capacity. Minor changes in the system could result in congestion and delays. One of the many impacting factors of transportation is weather condition. Weather conditions might lead to a totally different setting for management of transportation systems. Since weather is predictable, being able to measure the impact of weather conditions on transportation systems would help in better transportation management.

Estimating dependency of travel time on weather condition will enable us to predict more accurate travel time. But it is possible that not all components of weather impact travel time equally. There are several other factors associated with travel time that interact with weather conditions to affect travel time. Other questions raising from this are:

1. Which weather component impacts travel time the most?
2. Is the impact of weather on travel time a function of time?

The exercise investigates regression models to understand the effect of weather condition, accidents, and time on travel duration. Based on the identified factors parametric and non-parametric classifiers are implemented to provide class-based predictions. Lastly, the machine learning models are the rated based on accuracy, precision, recall, and Cohen Kappa score, and envisioned for various use cases.

Introduction

An intelligent transportation system is reliable and efficient. The reliability of transportation systems is measure of value of time. A critical component of such systems is, travel time prediction. High-quality automatic vehicle identification devices make it possible to perform short-term traffic flow analysis and develop forecasting techniques. However, the factors contributing to the unpredictability of traffic systems, including accidents, erratic driver behavior, and various weather conditions make predicting travel time very challenging (Qiao, Haghani, & Hamedi, 2012).

It is well recognized that the transportation system may be significantly disrupted by adverse weather events. Apart from extreme events like floods, tornados, and hurricanes that could be disastrous, milder changes in weather caused by rain, snow, and ice could also have apparent negative impacts on the system (Xu Zhang & Chen, 2019). With the increasing quality and quantity of near real time data and sophisticated analysis techniques, accuracy in predicting impact of weather conditions on travel time has increased significantly.

The aim of this study is to quantify the impact of different weather condition on travel time. The scope of the study is limited to parameter of demand side of transportation systems.

Management of transport systems both in terms of supply management and demand management is important for ensuring efficiency of urban transportation systems. Since most transportation systems operate at capacity, even small changes in demand or supply due to weather may significantly increase traffic delay that results in a totally different setting for management. Answering the central research question will help modify planning decisions and improve transportation system under various weather conditions.

Theoretical Background

Literature Review

Literature reviewed during the project is focused on two themes: travel time prediction models, and big data analytics. Prediction models are used in various areas of application

ranging from market behavior to population. According to (Tsapakis, Cheng, & Bolbol, 2013) prediction models for travel time prediction can be divided into two categories, parametric methods and non- parametric methods. Parametric models use statistical techniques like ARIMA, SARIMA, linear regressions that try to understand the data based on specified parameters. Whereas nonparametric methods include K nearest neighbor and random forest like technique that deal with the state of the entirety of the data rather than the specified parameters. There are advantage and disadvantages associated with both prediction method types.

The basic practice of linear regression suggests making sure that all the independent variables must be statistically independent. This is a challenging situation since most weather conditions and traffic characteristics like density, peak hour is highly correlated.

ARIMA models employ the internal relationships obtained from historic data; however, large variations in the historic data set would generate significant prediction error. In addition, most traffic systems exhibit nonlinear relationships, which makes it difficult for linear models to capture stochastic characteristics.

The basic K Nearest neighbor model is based on historic database, it looks at neighborhood similarity to find nearest neighbors in a continuous interval. It is required that the historic data be large enough for the model performance. Nonparametric models require considerably large processing time and storage space.

The Highway Capacity Manual states that in light rain, a 1.9 km/h reduction in speed during free-flow conditions is typical (Council, 2000).

(Xiaoyan Zhang & Rice, 2003) uses linear model for short term travel time prediction. They suggest that

According to (Goodwin) aggregate weather effects accounted for roughly 12% of travel time delay. In the same study he conducted a detailed analysis in Washington metropolitan area average delay increased by 21% on days with adverse weather.

In (Xu Zhang & Chen, 2019) the authors use quantile regression with dummy variables for rain and snowfall occurrence to observe the relationship. The suggested that response to

weather condition is not an average phenomenon hence to capture the relationship they use quantile regression as it is sensitive to the distribution. This lets them explore difference in impact of weather conditions on travel time across the distribution. The study adapts a decision tree-based classification technique.

(Tsapakis et al., 2013) explores the impact of weather condition on travel time on a macro level. In the study authors explore the average delay in travel time due to weather conditions on a city level using simple weighted and aggregate methods. They observed that the ranges of the total travel time increase due to light, moderate and heavy rain are: 0.1–2.1%, 1.5–3.8%, and 4.0–6.0% respectively. Light snow results in travel time increases of 5.5–7.6%, whilst heavy snow causes the highest percentage delays spanning from 7.4% to 11.4%. Temperature has nearly negligible effects on travel times. The study also mentioned that the impact of weather varied based on locational attributes.

In (Nookala, 2006) the author first performs correlation coefficient analysis to understand which weather parameter affects the traffic, daily traffic volume variability under different weather conditions to study on the influence on trip demands, congestion analysis to gauge how severe the weather impact is on traffic. Using these observations, he finally develops a time varying regression model to account for the impacts observed in the last step.

Conceptual Framework

Conceptual framework developed for the study investigates the possible factors that can impact travel time. Weather conditions do not impact travel time directly. As shown in table 1, weather conditions affect road environment and transportation management which in turn then impacts travel time. User characteristic include type of vehicle, day of the weak, purpose of the trip and drivers' characteristics. These factors directly impact travel time. It has also been observed in previous studies that drivers' characteristics are also affected by weather condition. The most affected driver characteristic is purpose of the trip. In previous studies it was observed that delays in the trip unsystematic errors refer to accidents or other

unpredictable occurrences. Road environment and transportation system performance are the big envelopes that cover all the characteristic of road travel.

Table 1 (Goodwin) lists the direct impact on demand and supply side of transportation operations due to different weather condition.

Table 1: Causal chain

Weather Events	*Roadway Environment Impacts*	*Transportation System Impacts*
Rain, Snow, Sleet, Hail & Flooding	– Reduced visibility – Reduced pavement friction – Lane obstruction & submersion – Increased chemical and abrasive use for snow and ice control – Infrastructure damage	– Reduced roadway capacity – Reduced speeds & increased delay – Increased speed variability – Increased accident risk – Road/bridge restrictions & closures – Loss of communications/power services – Increased maintenance & operations costs
High Winds	– Reduced visibility due to blowing snow or dust – Lane obstruction due to windblown debris & drifting snow	– Increased delay – Reduced traffic speeds – Road/bridge restrictions & closures
Fog, Smog, Smoke & Glare	– Reduced visibility	– Reduced speeds & increased delay – Increased speed variability – Increased accident risk – Road/bridge restrictions & closures
Extreme Temperatures & Lightning	– Increased wildfire risk – Infrastructure damage	– Traffic control device failure – Loss of communications & power services – Increased maintenance & operations costs

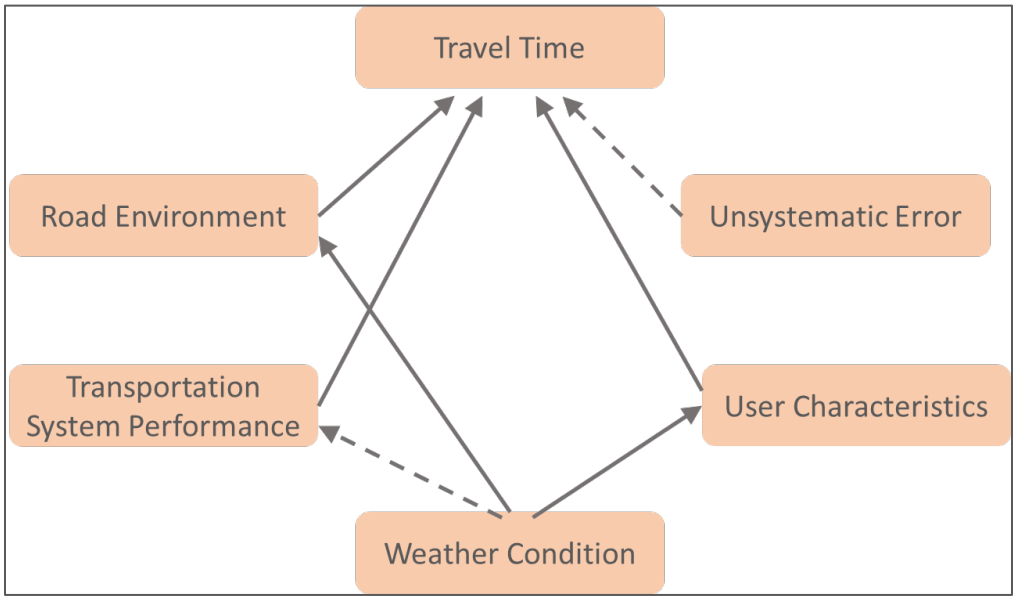


Figure 1: Conceptual framework to explain causal relationship between weather condition and travel time.

Based on the conceptual framework and previous work on travel time prediction various regression models were explored to find the nearest model. Model variables are shown in Figure 2**.** These variables are selected based on the data used for the exercise that is discussed in the later chapters and the research objective. Our dependent variable is travel time. Weather conditions are further divided into different elements to understand impact of each separately. Peak hours, day of the week, number of left turns are included in the model to improve the accuracy of the model. Accident data is also attached to the model as a dummy variable to account for delay due to them. Since the data used for the exercise is collected only for yellow rides it eliminates any systematic variables related to vehicle characteristics. Other driver characteristic is not addressed in the exercise due to limitation of the data. The project also explores relation of parts of the city and travel time. Development of these clusters is explained in the later parts.

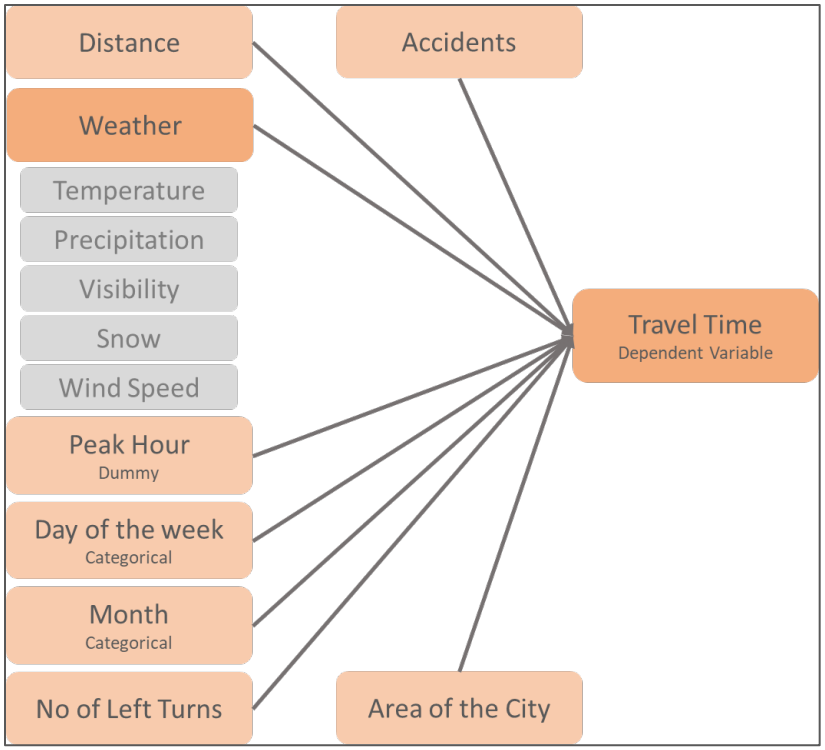


Figure 2: Model Variables

Data Requirements

There are three primary data sources used in the project trips data, weather data and accident data, apart from that various other structural data are used like open street maps and New York city demographic data.

Cab rides for 2016 were selected as they had precise latitude and longitude of pickup and drop off locations. After 2016 the data available only had coordinates for the origin and destination taxi zones. The data used from 1st January 2016 to 30th June 2016. This contains a total of 1,458,643 rides. The average duration of a trip is 959.5 seconds**.** Distribution of the number of trips over the 6 months is shown in figure 3.

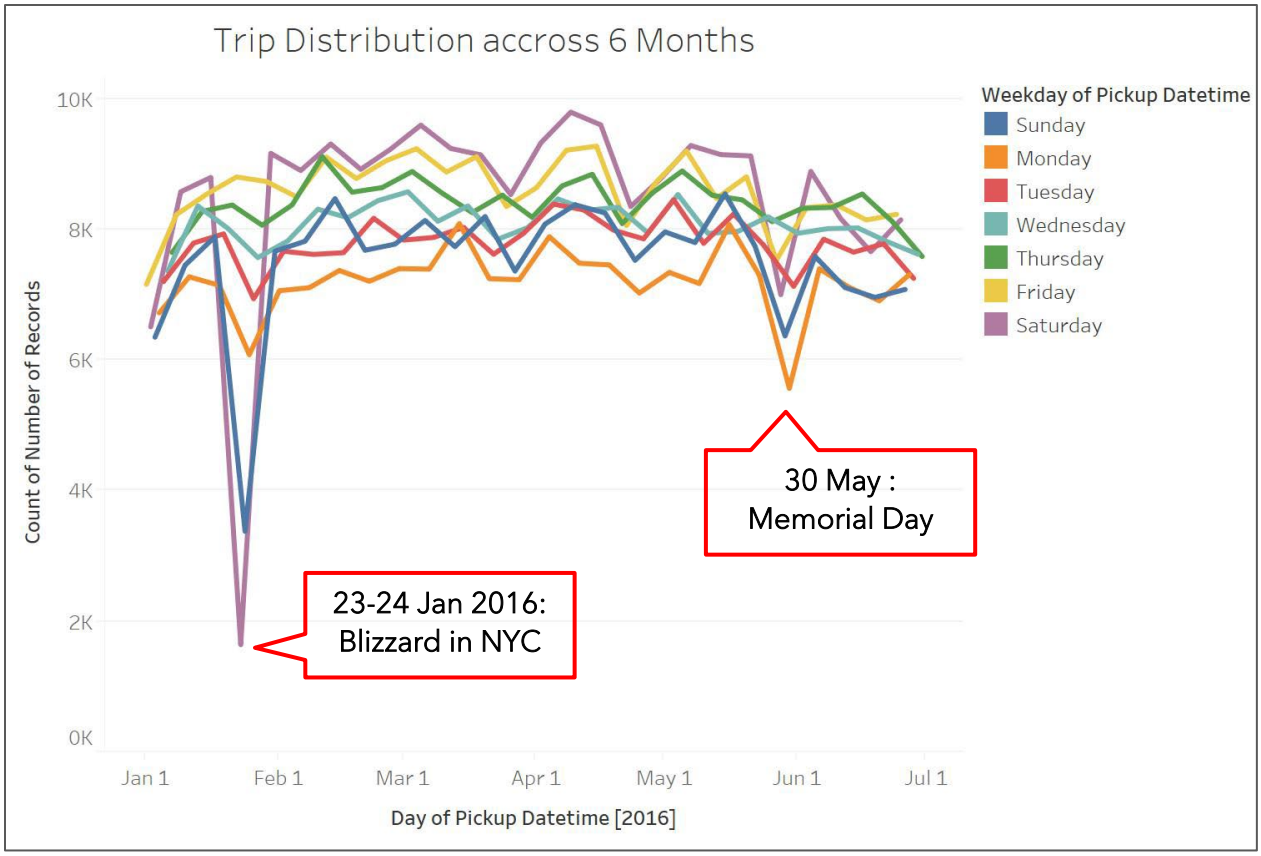


Figure3: Trips Distribution across 6 months

Weather data is collected from Integrated Surface Dataset (ISD) , NOAA**.** Based on the extent of the study are 4 weather stations were selected. The data contains name and ID of weather stations, timestamp of recorded value and various weather attributes. Most weather conditions were recorded at hourly and a few were collected at daily frequencies.

Accident data is collected from NYC Open Data**.** Accident data account for the recorded vehicle collision that occurred in the city during the time. There was a total of collisions in the city. Out of which 97,834 are used due to limitations in the data quality and accuracy.

Table 2: Data Source

Data	Source
Cab Rides	TLC Trip Record Data (https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page)
Weather Data	Integrated Surface Dataset (ISD) , NOAA (https://gis.ncdc.noaa.gov/maps/ncei/cdo/hourly)
Accident Data	Motor Vehicle Collisions, NYC Open Data (https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/data)

Methodology

Since the project deals with a high volume of data (over 3 GB) a lot of efforts have been put to design the methodology and choosing tools that support big data analytics. The methodology devised for the project is divided into five parts:

i. Data Cleaning
ii. Data Preparation
iii. Analysis
iv. Data Integration and
v. Prediction Model

Data preparation included data collection, cleaning, and generation of derived variables that are but not limited to calculating route distance, calculating number of left turns, interpolation rules for weather condition, understanding peak hours, etc. Data Integration involves creating rules to stitch all the variables together.

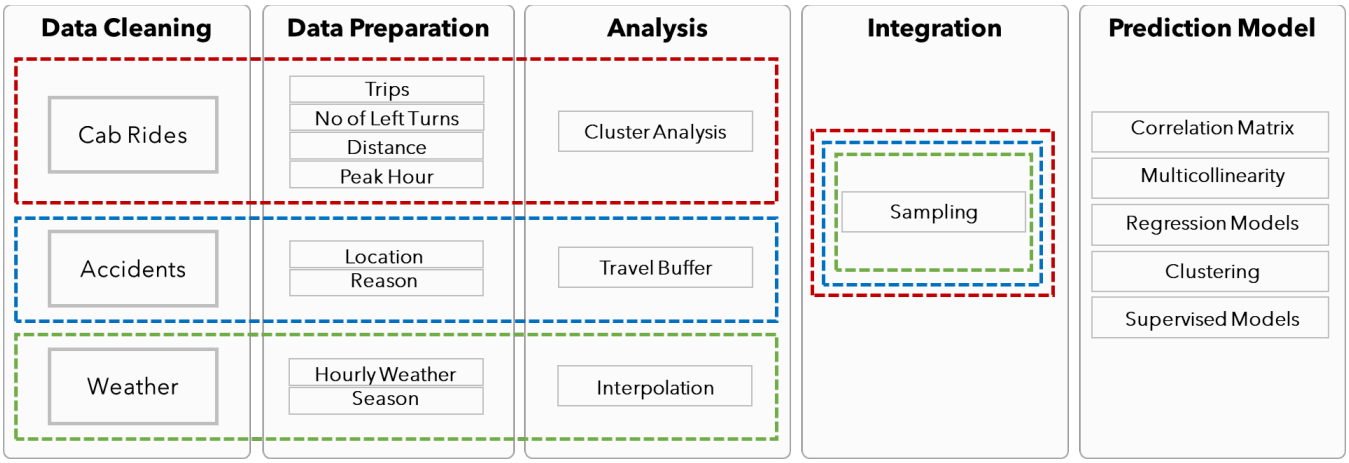


Figure 4: Methodology

Multivariate modeling part of the methodology explores various regression models to find the best possible model for understanding relation between independent and dependent variables. We start with sampling methods that need to be adapted to handle big data. The next step is to scale all the variables on same scale using Z- transformation. Z-transformation makes sure that all the variables are comparable. The next step for running regression model is to get rid of collinearity between our variables. To measure this collinearity matrix is developed. Based on this matrix highly correlated parameters are removed. Finally, various regression models are observed. Each step is explained in detail in the next chapters.

Data Cleaning and Preparation

Dask module is used to first stitch all the cab rides together, collected for 6 months starting from Jan 2016 to June 2016. Once all the cab rides have a homogeneous structure this data frame is then switched to PySpark data frame. This step was necessary as Dask is a relatively new module, finding documentation and test case handling of the module is relatively poor. In the next step data types were assigned to all the column. The next step was to get rid of nan values from the data and clipping out cab rides going outside or coming from outside the city boundary.

The next step was to calculate other related derived variables like distance, time elapsed and number of left turns for each cab ride. Duration of the trip is simply calculated based on the

timestamps of the rides. To calculate the distance between the pickup and drop off point it is assumed that the trip would take the shortest possible path. This distance is calculated using Open Source Routing Machine (OSRM) API. OSRM as the name suggest is an open source platform that performs routing-based calculation for the input data and gives the data in as a json object. To calculate the distance and number of left turns in the route taken, coordinates for origin and destinations are passed through the api. This API the gives result as a json object, finally using designed functions distance and left turns are calculated. An example of this JSON object is shown in figure 5.


```
{
 "code": "Ok",
 "waypoints": [
  {
   "hint": "1T2WgP___381AAAARgAAAAoAAAAJAAAAYq8-Qsc0aEEy1RRBMtUUQ
TUAAABGAAAACgAAAAkAAAAxTwAAUSeX-3mwbQIqJ5f7irBtAgEA3wEmxxW3",
   "location": [
    -73.980079,
    40.743033
   ],
   "name": "3rd Avenue"
  },
  {
   "hint": "voSThcOEk4UBAAAAAAAAAGcAAAAKAAAAXelwPwAAAAC1Yo5CIr7tQ
AEAAAAAAAAAZwAAAAoAAAAxTwAAiiuY-zD_bQJuK5j7Bv9tAgIA7wcmxxW3",
   "location": [
    -73.913462,
    40.763184
   ],
   "name": "30th Avenue"
  }
 ],
 "routes": [
  {
   "legs": [
    {
     "steps": [
      {
       "intersections": [
        {
         "out": 0,
         "entry": [
          true
```


Figure 5: OSRM json output

Weather data collected has hourly reading for precipitation, temperature, visibility and daily averages for snowfall and snow depth. The data contains IDs of the weather stations. Coordinates od these stations are first added. All the data processing for this dataset has been performed in pandas. The first step was to create hourly weather condition from daily averages for snow fall and snow depth. It was assumed that these values would remain

constant through the day. It was observed that the data obtained was not strictly hourly, some hourly rating were collected multiple times withing one hour. In order to get rid of multiple values within the same hour, data frames were grouped on hourly basis of each weather station. Using designed functions on these hourly data frames we got rid of all the nan values, multiple values, and missing data points. In case of nan values in all the points within the hour 0 was replaced. For hours with multiple values average value was taken. In cases with nan and other values for the same hour, nan was removed, and the other value was used. The function was designed in a way to differentiate between 0 and nan values in order incorporate all the cases as discussed above. This function was used to generate hourly data points for each weather condition at each weather station. Finally, the generated values were compiled to get the desired weather dataset.

Finally, accident data is also processed using Pandas data frame. Accident data contained time of the accident, coordinated of the accident, street name on which the accident occurred, number of people injured and cause of accident. It was observed that some of the data points did not contain any locational attribute. It was impossible to generate that information using this data sets hence such points were removed firstly. Only about 20% of the data contained precise coordinates of the accident, to get location of rest of the datapoint street names were used. Using Google's geolocation API coordinates for the rest of the points were generated. Finally, locations lying outside the study are were cropped and the remining points were plotted. The most vulnerable areas are highlighted using point kernel density function (figure 6).

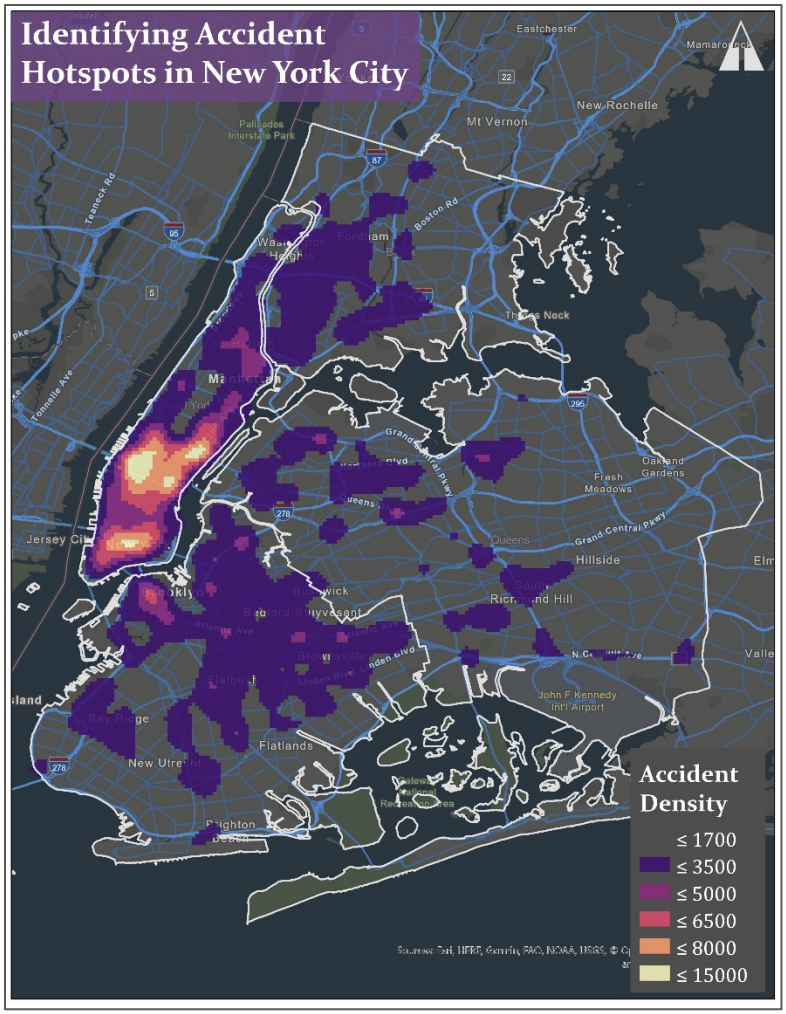


Figure 6: Accident Density

Analysis and Integration

Using the filtered dataset with derived variables we will now explore other analysis that are required before developing the regression model. To understand how these trips, connect and different parts of the city a cluster analysis was performed on small sample of trips. Using random sampling 20% of the data was used to achieve this. The pickup and drop off location of these point were snapped to the underlying census block. This was necessary to develop a network encompassing all parts of the city. Using this block to block network cluster analysis was performed in R using i-graph package. The cluster analysis assigns modularity classes or group number to each node which in this case is center of

census block. These clusters are modularity classes generated based on network connectivity. As shown in Figure 7 the clusters are spatially contiguous. These clusters will be used as categorical variables in the regression model to check if the impact of weather on travel time also a function of area of the city.

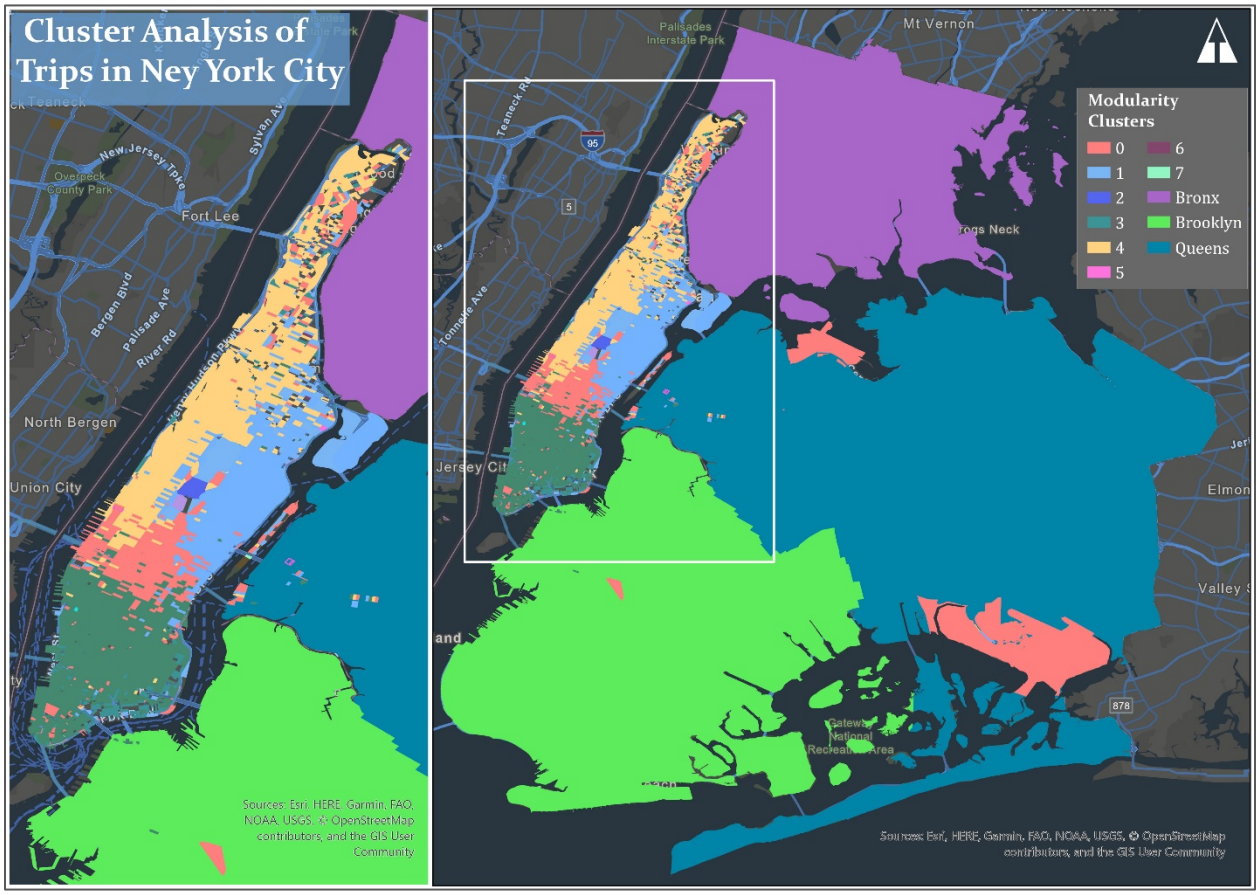


Figure 7: Cluster Analysis of Trips

Weather conditions at stations need to be interpolated though the whole study area. As previously stated, there are 4 weather stations, for each station there are 5 weather attributes, and for each weather condition there are hourly readings collected for 6 months. This is not a feasible solution. Hence to avoid generating so many interpolations surfaces the project uses various conditional function to optimize this process. The conditions are employed on each weather condition separately:

- If the difference in weather conditions at 4 weather station is less than a specific value (different for each condition), then take average value for that hour

- If the difference in weather conditions at 4 stations is more than this specific value, interpolate to surface raster.

To account for impact of accidents on trips, a travel buffer of 200m is applied around the location of accident along the streets. This buffer is applied as an intent to account all the trips that will pass around the area and get delayed due to the collision. Accidents are added as dummy (1,0) to trips if the time of accident in within 15 min of trip end time.

The final step is to compile all our variables into one data frame. The integration process is completed using various pipelines. The base data for integration is trips. The final product is trips with all the variables added based on timestamps.

Prediction Model

Statistical Analysis

Figure 8 compares distribution of number of trips in a day. The identified peak from the distribution are Morning peak – 7 to 11 AM, Evening peak – 5 to 9 PM. Number of trips reached global minima around 5 AM. Figure 9 compares the average ideal travel time for the trips with the actual time taken. It shows that as the number of trips decreases and reaches global minima at 5 AM, the actual time taken is less than the ideal time.

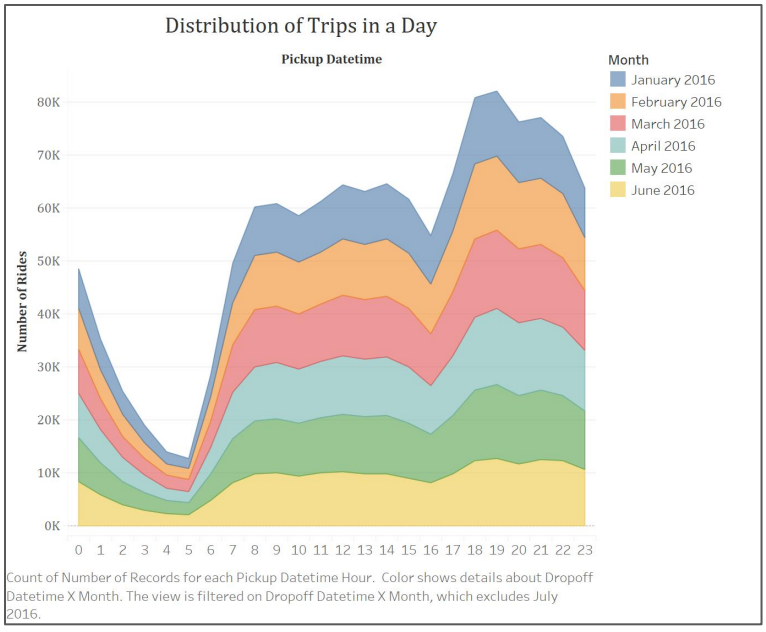


Figure 8: Distribution of Trips in a Day

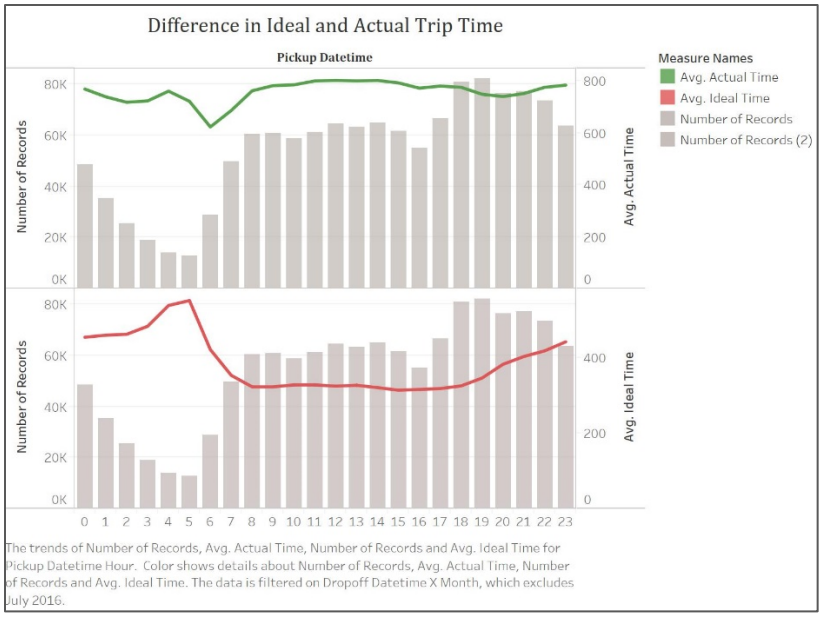


Figure 9: Difference in Ideal and Actual Time

The difference between two increases as the number of trips in the hour increases. Figure 10 shows the frequency of actual time. Data points outside 2 standard deviations have been removed to get rid of any possible outliers. As the distribution is not normal, during the sampling process data points are upscaled to make the spread normal.

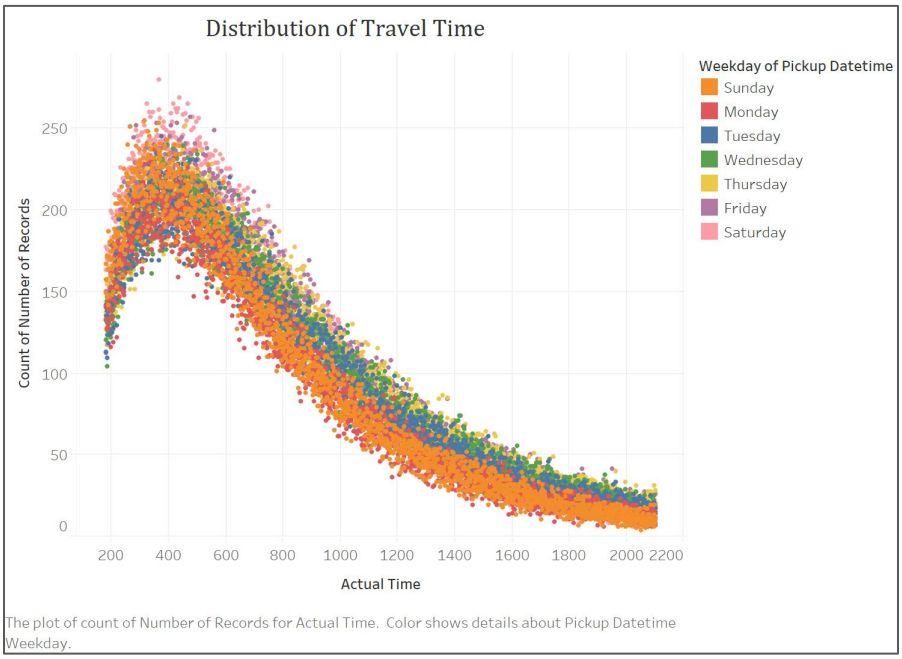


Figure 10: Distribution of Travel Time

Linear Regression

Before using linear regression models, collinearity between the variables were calculated. Firstly, correlation matrix was generated between all the variables. Figure 11 shows the output of the matrix. None of the pairs had correlation value greater than 0.5. Hence for the first linear model all the variables were used. The model gave the initial R – square. The model also had high p value.

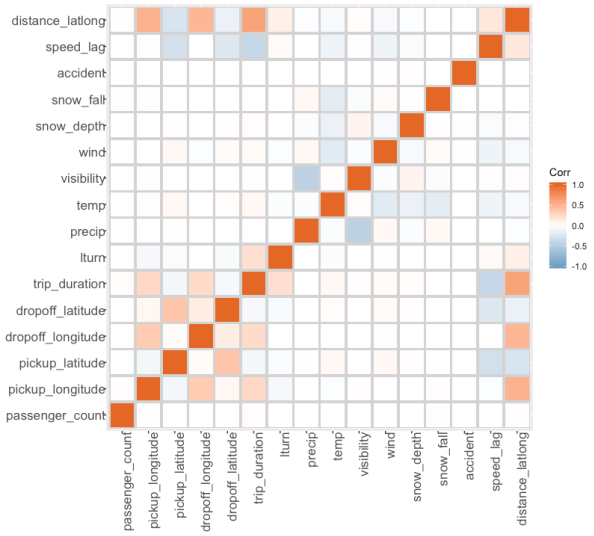


Figure 11: Correlation Matrix

The next step was to conduct Variance Inflation Factor analysis (VIF). VIF check for multicollineary between the selected variable and all other variables combined. The results showed high VIF(>5)for pickup and dropoff coordinates, for weather conditions and distance.

Based on VIF results the finalized variables for linear regression models were:

Dependent Variable: Travel time

Independent Variables:

- Direct distance from pickup to dropoff- since latitude and longitude were not a suitable parameter, I calculated the direct distance using the coordinates to get rid of the multicollinearity.

- Number of Left Turns
- Weather conditions: Snow Depth, Visibility, Temperature
- [As Dummy] Hour of the Day, Day of the Week, Month, Accident, Vendor

Linear regression model with these variables gave an R – squared value of 0.49. which mean that the model was able to account for 49% variance in the data increases the R square by 35% compared the first model.

The base case for the final regression model was Sunday, January, at 00:00 hour, with no accident and traveling with vendor 1. Base case sets the comparative case and all the calculation are done in comparison to this case.

Figure 12 shows the value of coefficients of the finalized linear regression model. It shows that there is a positive coefficient for snow depth (12.6) and temperature (0.2). That means as snow depth or temperature increases travel time also increases. There is negative coefficient for visibility (-6), hence an increase in visibility decreases travel time. The most significant variable was hour of the day, going as high as 248 to as low as -151. Occurrence of accident has coefficient of 36, which mean an accident on the trip could result in a delay of 36 seconds, keeping all other variable constant. Visit **Annexure** for other coefficients.

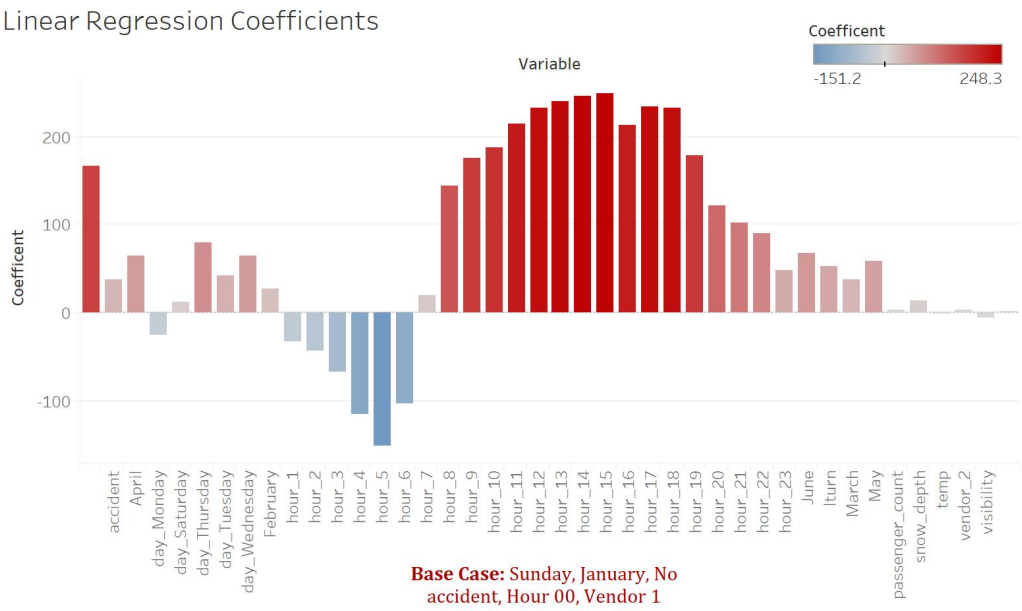


Figure 12: Linear Regression Coefficient

Stratified Sampling

It is clear from the visualizations that the distribution of the population set in terms of the time taken to complete the trip is not homogeneous. There is a 40-40-20 split among **extremely delayed** (cases where delay $>$ 5 meter/sec), **delayed** (cases where delay $<$ 5 meter/sec), and **not delayed** classes, which was computed through difference in ideal speed and actual speed through data and the OSRM output.

The most effective method for sampling when the population is heterogenous (suffering from class imbalance) is Stratified Sampling (htt)

The main idea behind the method is:

- Divide the heterogenous population into sub-groups (in the study, this is done on the basis of delay class), such that the units are homogenous with respect to the characteristic that is being studied
- The population data is heterogenous with respect to sub-population, or strata
- Each strata is treated as a separate population and draw a sample from each stratum with the same probability

N: Population size

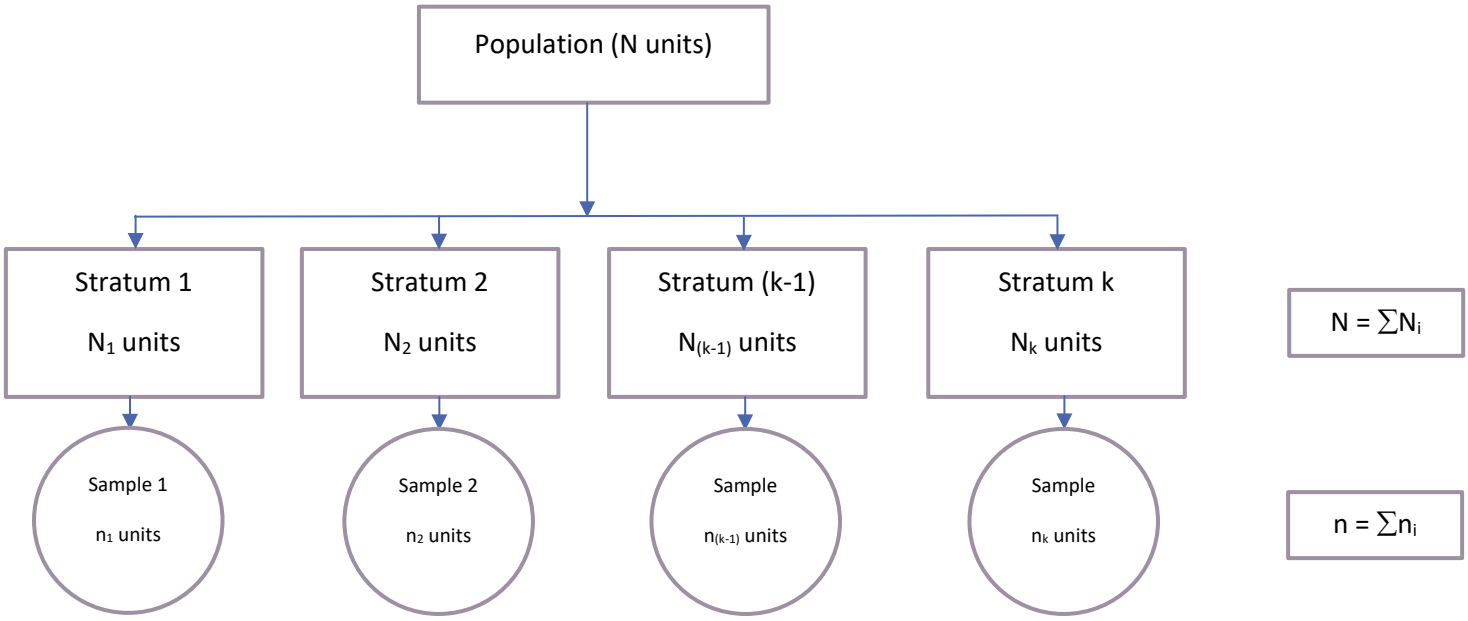
k: Number of strata

N_i: Number of sampling units in the i^{th} stratum

$N = \Sigma N_i$

ni: Number of sampling units to be drawn from the i^{th} stratum

$n = \Sigma n_i$



Stratified sampling ensures that the probability with which an element is picked from a stratum has the same probability of picking as that of any other stratum. Since the dataset contains 3 classes, the probability of picking an element was 0.33 for each class.

The classes in the original dataset, which were in the 40-40-20 proportion were now 34-33-33, which means the "no delay" class was up-sampled to match the number of instances in the other two classes.

Dimensionality Reduction

To apply feature engineering and machine learning techniques to conduct analysis and to build the predictive classification models, the number of dimensions that were present in the final regression model might requirement of calculations which spans through multiple dimensions, increasing complexity. Through the introduction of dummy variables for hour (23 dimensions), day (6 dimensions), month (5 dimensions), and vendor (1 dimension), the number of dimensions increased from 20 to 55 (only 45 used in the regression model).

It was important to check how the variance was distributed across various dimensions in the feature space. If the essence of the model could be encoded in less than 45 dimensions, it would greatly reduce the amount of computational power.

To verify if the exercise was worthwhile, it was necessary to investigate the computational complexity of the algorithms that are explored in the analysis (The Kernel Trip, n.d.) Considering "n" number of training samples, "p" number of features, "n_{trees}" number of

trees (for tree-based methods), "n_{sv}" number of support vectors, and t=min(n,p), following are the approximate computational complexities:

Algorithm	Training	Prediction
Principal Components Analysis	$O(p^2n + p^3)$	---
Linear Discriminant Analysis	$O(pnt + t^3)$	---
Linear Regression	$O(p^2n + p^3)$	$O(p)$
K-Nearest Neighbors	---	$O(np)$
Decision Tree	$O(n^2p)$	$O(p)$
Random Forest	$O(n^2pn_{trees})$	$O(pn_{trees})$
Stochastic Gradient Descent	$O(n^2pn_{trees})$	$O(pn_{trees})$
Support Vector Machine	$O(n^2p + n3)$	$O(p)$
Extreme Gradient Boosted Trees	$O(npn_{trees})$	$O(npn_{trees})$

As "n" and "p" play a huge role in various training and testing phases, it would be great for the classifiers to reduce the value of both "n" and "p", which is why sampling and dimensionality reduction was conducted.

Two methods were compared for performance of dimensionality reduction process:

- **Principal Components Analysis (PCA):**

  The various Principal Components (PCs) for the data were studied, and through Singular Value Decomposition (SVD) of the design matrix done through calculation of covariance matrix, and through Eigenvalue decomposition of the covariance matrix.

It was identified that 85% of the variance in the data could be explained through the first two PCs. The plot for the factor loadings w.r.t. the first two PCs:

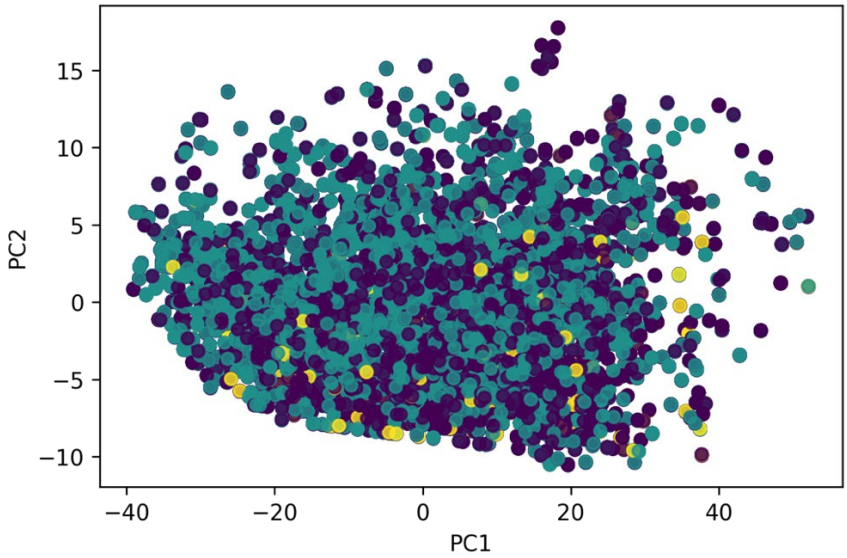


Figure 12: Principal Component Analysis Results

The color of each data-point represents the class that it belongs to, and PCA does not seem to identify clear separation bounds because it hardly gives any importance to the class label. Therefore, it is not very effective for our problem.

- **Linear Discriminant Analysis (LDA):**

LDA is a dimensionality reduction algorithm similar to PCA. However, it is different from PCA due to the importance LDA gives to classes. The algorithm behind LDA tries to increase the separation bounds between each class, and thereby identify optimal Linear Discriminants (LDs). It was identified that the first two LDs explained 93% of the variance in the data. The plot for the factor loadings for w.r.t. the first two LDs:

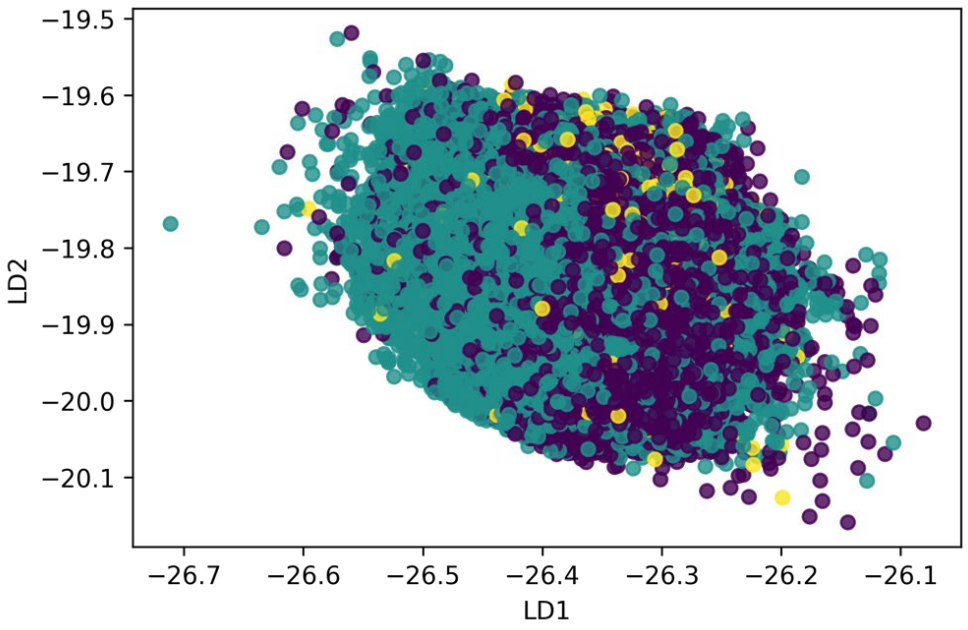


Figure 13: Linear Discriminant Analysis Result

As it is visible from the plot of LD1 vs LD2, the separation bounds among various classes are clearly visible through LDA, which was not demonstrated through PCA. The further steps involve using LD1 and LD2 as the primary dimension for the dataset, and we can discard the original form of the dataset.

For performing predictions, each of the test instance will first be projected on the LD1 and LD2 dimension, and the prediction would be made only on the basis of these new dimensions.

The next section discusses about various Machine Learning algorithms that were implemented for the dimensionally reduced dataset.

Machine Learning

- **Unsupervised Learning:** The k-means clustering algorithm was implemented for the range $1 < k < 10$ to find the optimal value of k. The elbow plot for the experiment is as follows:

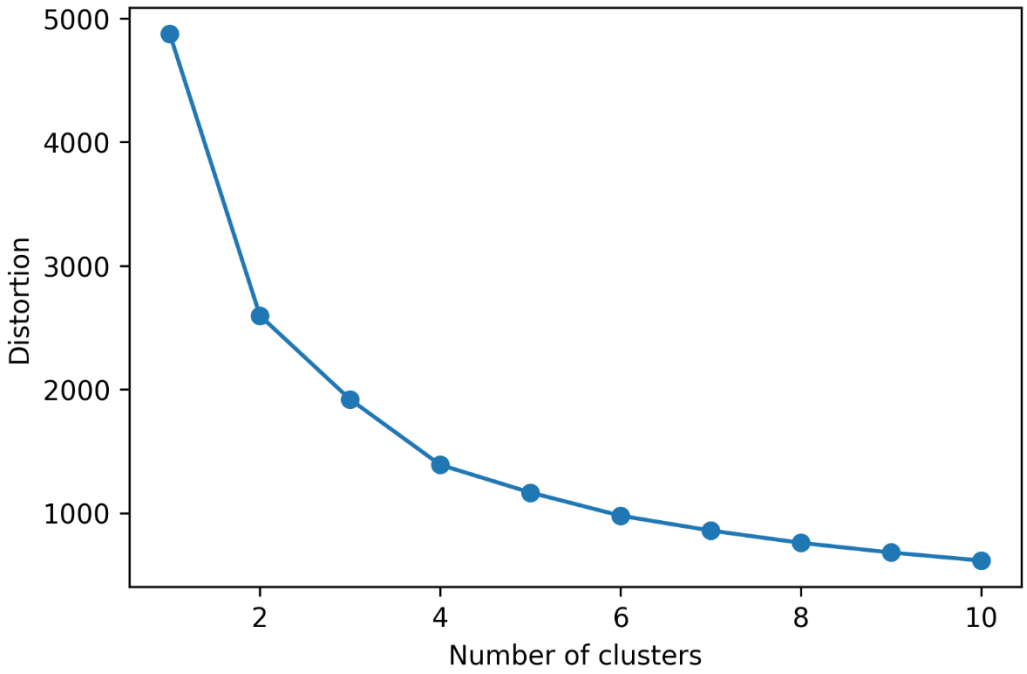


Figure 14: Distortion Vs no of clusters

As it is visible through the elbow plot, the value of k=3 demonstrates a sharp decrease in the distortion and is the location of the elbow. This means, the optimal number of clusters for the problem is 3, which in extension suggests that 3 classes for predictive modeling is a good estimate, and the supervised algorithms that are implemented below are in concurrence to the results of the unsupervised learning experiment.

- **Supervised Learning:** Both LDA and K-means clustering have been used in supervised and unsupervised learning due to their ability to consider class labels to perform prediction. The performance of various learning-based classifiers are listed below:

Model Name	Accuracy	Cohen Kappa Score	Precision			Recall			F1-Score		
			Delay	Extreme delay	No delay	Delay	Extreme delay	No delay	Delay	Extreme delay	No delay
Linear Discriminant Analysis (LDA)	60.10%	0.39860	50.15%	64.79%	63.90%	47.63%	75.13%	55.70%	48.86%	69.58%	59.52%
Decision Tree Classifier	65.48%	0.48196	53.82%	63.83%	75.54%	47.24%	60.27%	89.45%	50.31%	62.00%	81.91%
K-Means clustering	59.15%	0.38586	47.15%	67.33%	67.13%	59.44%	70.72%	45.93%	52.59%	68.98%	54.54%
Support Vector Machine	60.91%	0.41160	51.25%	66.62%	62.71%	46.64%	72.57%	62.06%	48.84%	69.47%	62.39%
k-Nearest Neighbour (k=7) Classifier	60.17%	0.40164	50.07%	65.85%	62.70%	45.40%	66.05%	68.31%	47.62%	65.95%	65.39%
Gaussian Naive Bayes Classifier	59.60%	0.39035	49.62%	63.50%	64.54%	46.67%	77.20%	52.78%	48.10%	69.69%	58.07%
Random Forest Classifier	68.13%	0.52157	57.00%	67.66%	77.24%	51.32%	64.68%	88.70%	54.02%	66.13%	82.58%

Stochastic Gradient Descent (SGD) Classifier	55.47%	0.32296	57.00%	67.66%	77.24%	51.32%	64.68%	88.70%	54.02%	66.13%	82.58%
Gradient Boosting Classifier	61.04%	0.41400	51.50%	67.48%	61.96%	46.42%	71.30%	64.12%	48.83%	69.34%	63.02%

Conclusion

Impact of weather condition on transportation safety, mobility and production is inevitable. However, these impacts can my significantly reduced through smart transportation policies. According to (Goodwin) there are three types of mitigation measures that can be adapted to reduce this impact, control, treatment and advisory strategies. They define control strategies as devices to permit or restrict traffic flow and regulate roadway capacity. Treatment strategies are using resources to reduce the impact of weather. And advisory strategies deal with providing information on weather conditions, it serves in supporting decision making process.

Management strategies require relevant, accurate, and timely environmental data to effectively mitigate weather effects. Managers need observations and predictions of road weather conditions to make operational decisions (Goodwin). These mitigation strategies are adapted based on the scale of expected impact of weather condition. Scale of decisions taken to mitigate such situations can be planning, warring or Operations based. A robust prediction model for quantifying weather impact will help in developing better road weather management strategies.

Works Cited

- (n.d.). Retrieved from http://home.iitk.ac.in/~shalab/sampling/chapter4-sampling-stratified-sampling.pdf
- (n.d.). Retrieved from The Kernel Trip: https://www.thekerneltrip.com/machine/learning/computational-complexity-learning-algorithms/
- Goodwin, P. P. a. L. C. *Surface Transportation Weather Applications*.
- Nookala, L. S. (2006). *Weather Impact on Traffic Conditions and Travel Time Prediction.*
- Qiao, W., Haghani, A., & Hamedi, M. (2012). Short-Term Travel Time Prediction considering the Effects of Weather. *Transportation Research Record: Journal of the Transportation Research Board, 2308*(1), 61-72. doi:10.3141/2308-07
- Tsapakis, I., Cheng, T., & Bolbol, A. (2013). Impact of weather conditions on macroscopic urban travel times. *Journal of Transport Geography, 28*, 204-211. doi:10.1016/j.jtrangeo.2012.11.003
- Zhang, X., & Chen, M. (2019). Quantifying the Impact of Weather Events on Travel Time and Reliability. *Journal of Advanced Transportation, 2019*, 1-9. doi:10.1155/2019/8203081
- Zhang, X., & Rice, J. A. (2003). Short-term travel time prediction. *Transportation Research Part C: Emerging Technologies, 11*(3-4), 187-210. doi:10.1016/s0968-090x(03)00026-3