BUSINESS ANALYTICS REPORT

**Analysis & Prediction ChicagoTaxiDemand**

### By



### Prathamesh

### Sawant

**TableofContents**

|  |  |
| --- | --- |
| **OVERVIEW** | **3** |
| **METHODOLOGY** | 6 |
| **EXPLORATORYDATAANALYSIS** | 8 |
| **DATAPRE-PROCESSING** | 14 |
| **MODELING** | 18 |
| **EVALUATION** | 24 |
| **REFLECTION** | 27 |
| **CONCLUSIONANDFUTURESCOPE** | 27 |
| **REFERENCES** | 28 |

# Overview

In this project, we analyze the taxi industry in Chicago, specifically the taxi tripsthattook place in the city of Chicago for two consecutive years - 2019 and 2020.Forthis project,wepulledthedatafromtheChicagoOpenDataPortalandusedR as our programming language. Using CRISP-DM methodology, we aim to answ

er the following questions:

1. What were the trends in the Taxi industry before the COVID-19 pandemic,and how did it change a year later?
2. Which Machine Learning algorithm will be best suited for building a modelto predict the taxi demand?

## BusinessOverview

The taxi industry in Chicago has a rich history, starting from 1853 when the firsttaxi company was found to cater to the transportation needs of people using therailway [1]. In the 1920s, the two biggest taxi companies in Chicago, Yellow CabCompany and Checker Taxi were engaged in a full-out rivalry which includedshooting in broad daylight and the taxi drivers engaging in tank warfare maneu-verers to beat the competition [2]. In recent years, the prevalence of ride-hailingcompanies like Uber has started to take over a significant share of the market,withUberclaimingtohaveprofited$46,380,000,andcreating25,000incrementalrides in 2013 [3].

TheCOVID-19pandemicwreakedhavoc onthetaxiindustry,withthetraditionaltaxi company hit the hardest, since they were already dealing with the blowbackfrom the emergence of rail-hailing industries. Considering the significance of thetaxi industry we will be analyzing the taxi trips reported by the cab companiesinChicago. Also, we will be analyzing data from 2019 and 2020, years envelopingthe pandemic, to better capture and report trends in the demand for taxis.

By building a reasonably accurate predictive model, we hope that it can be usedby the taxi companies to schedule their taxi fleet more efficiently, resulting in re­ducedpassengerwaitingtime,betterutilizationofthefleetresourcesandan

increase in incremental rides. We will be building the model using various ma­chine learning algorithms and comparing their performances.

## DatasetOverview

The city of Chicago hosts a publicly accessible datastore which contains around600datasetscontaininginformationoncitydepartments,publicservicesandtheirperformances, for the benefit of researchers. The taxi trip dataset is populatedusing information captured from the two biggest payment processors in serviceforthetaxi companies. Inthedataset, rides from ride-hailing companies suchasUberandLyftarenotbeingrecordedandthusnotinthescopeofthisproject.Thedataset is periodically updated by the city of Chicago and it contains around 198millionrecords.However,theobservationsduring2019and2020accountforap­proximately 20.8 million.

Thedatasetcontains23columns,being:

|  |  |  |
| --- | --- | --- |
| **SNO** | **COLUMNNAMES** | **DESCRIPTION** |
| 1 | **TripID** | UUIDforeachtrip |
| 2 | **TaxiID** | UUIDforeachtaxi |
| 3 | **TripStartTimestamp** | Tripstarttime(roundedtothenearest15minutes). |
| 4 | **TripEndTimestamp** | Tripendtime(roundedtothenearest15minutes). |
| 5 | **TripSeconds** | Thetripdurationisinseconds. |
| 6 | **TripMiles** | Thetotaldistancecoveredinmiles. |
| 7 | **PickupCensusTract** | TheCensusTractfromwherethetripbegan. |
| 8 | **Dropoff CensusTract** | TheCensusTractfromwherethetripended. |
| 9 | **PickupCommunityArea** | TheCommunityAreafromwherethetripbegan. |
| 10 | **DropoffCommunityArea** | TheCommunityAreafromwherethetripended. |
| 11 | **Fare** | Thefareforthetrip. |
| 12 | **Tips** | Thetipforthetrip.Cashtipsarenotrecorded. |
| 13 | **Tolls** | Thetollsforthetrip. |

|  |  |  |
| --- | --- | --- |
| 14 | **Extras** | Anyextrachargesincurredduringthetrip. |
| 15 | **TripTotal** | Totalcostofthetrip. |
| 16 | **PaymentType** | Typeofpaymentused. |
| 17 | **Company** | Thecompanyunderwhichthetaxiisregisteredfor. |
| 18 | **PickupCentroidLatitude** | Thelatitudeofthecenterofthepickupcensustractorthecommu­nity area if the census tract has been hidden for privacy. |
| 19 | **PickupCentroidLongitude** | Thelongitude of the center ofthe pickup census tract or the com­munity area if the census tract has been hidden for privacy. |
| 20 | **PickupCentroidLocation** | Tupleofpickupgeo-coordinates. |
| 21 | **DropoffCentroidLatitude** | The latitude of the center of the drop-off census tract or the com­munity area if the census tract has been hidden for |
| 22 | **DropoffCentroidLongitude** | Thelongitudeofthecenterofthedrop-offcensustractorthecom­munity area if the census tract has been hidden for privacy. |
| 23 | **DropoffCentroidLocation** | Tupleofdrop-offgeo-coordinates |

Tomaintainprivacyandensurethatthispublicdataisnotbeingexploitedformalicious use, certain provisions have been taken:

1. Thetripstartandendtimestampshavebeenroundedtothenearest15minutes.
2. ThetaxilicensenumberismaskedusingaUUID(UniversallyUniqueIden­tifier).
3. Census Tracts having less than 3 trips in the relevant 15-minute time slotare not shown.

Otheroutliersareautomaticallynotupdatedinthedataset,suchas:

1. Triptimeslessthan0andmorethan86,400seconds.
2. TripMileslessthan0andmorethan3500miles.
3. Tripcostlessthan$0andmorethan$10,000.

# Methodology

Ourprojectisbroadlydividedintotwomainobjectives:

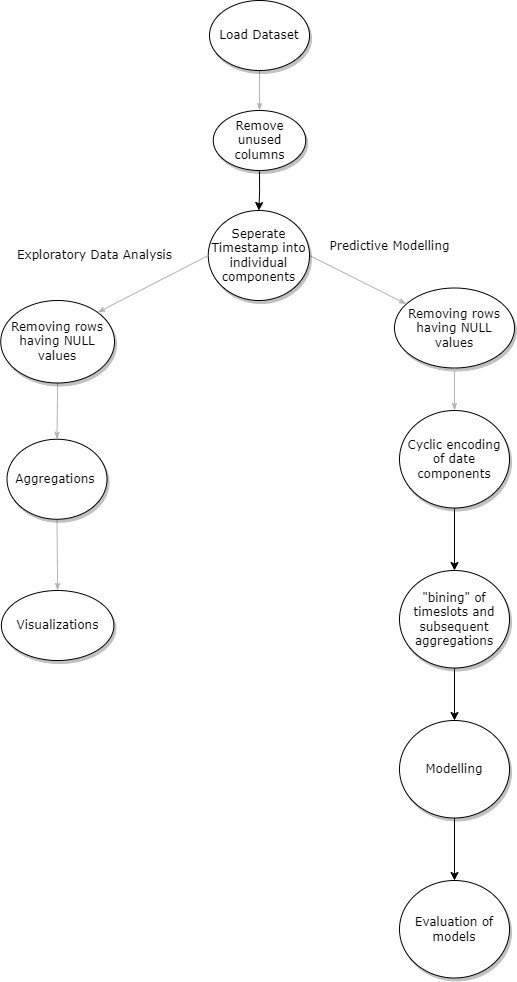
1. Comparingpre-covidandpost-covidtrendsfortheChicagotaxiindustryus­ing Exploratory Data Analysis.
2. Use a set of preselected Machine Learning algorithms to build a model forpredicting taxi demand and compare their performance.

Fortheeffectiveperformanceofourmodels,weareonlyusingthedatafrom2020so that our models can understand the trends in the covid scenario better.

Since we are processing a large dataset, we ensured that proper error handlingis in place, including batch processing of the original dataset.

For the modeling part, we split the pre-processed dataset into 70% training da­tasetand30%testingdataset.ThedatasetsplittingwasNOTrandomizedtopre­serve the trend over time.

The figure below depicts the flowchart of the various processes involved inourproject.



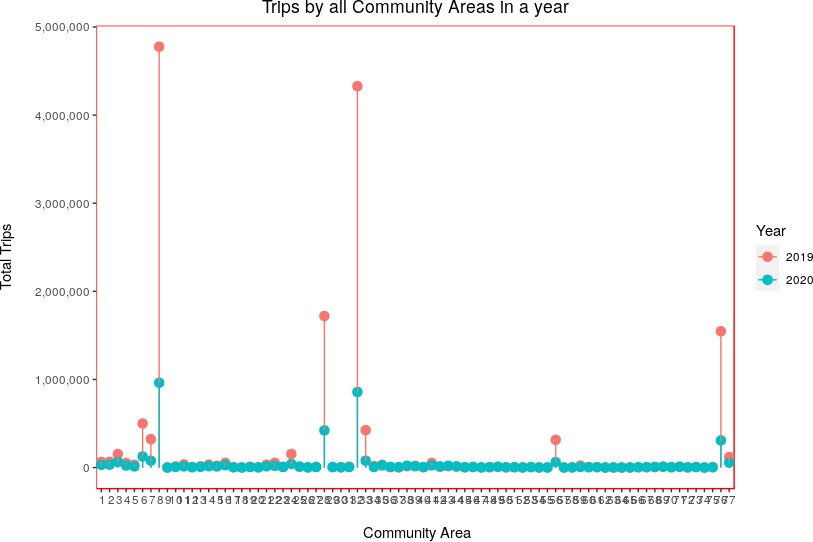
# ExploratoryDataAnalysis

As a first step, we explored the Chicago Taxi dataset and we had our observa­tionsrecorded.Wecomparedvariousattributesofthedataandanalyzedthere­lationships between them for 2019 as well as 2020.

Wehaverecordedourobservationsthroughgraphsandinferredthevariouspatterns reflected in the existing data.

##### TaxiTripsbyCommunityAreasinChicago

ThecityofChicagoisdividedinto77communityareas.Thebelowgraphshowsthe number of taxi trips availed in a community area before and after Covid.

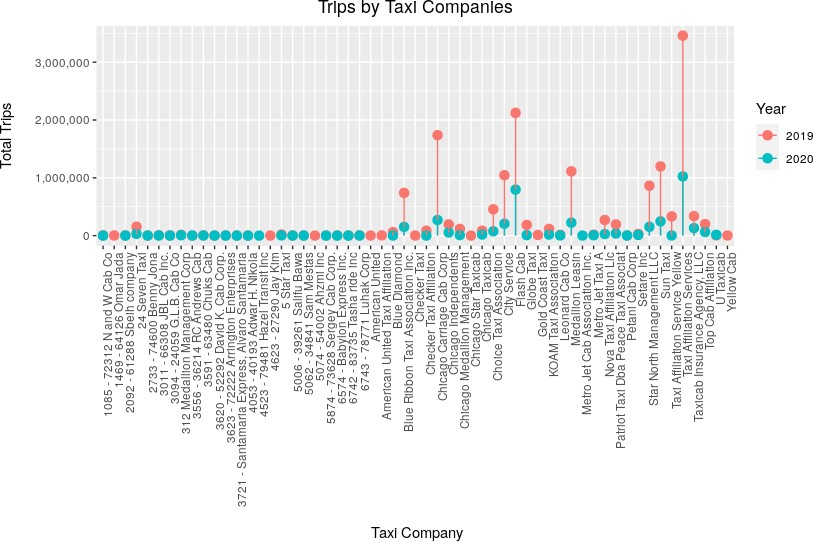


From the data we can infer that the following community areas have the highestnumber of taxi trips availed before and after the pandemic:

|  |  |  |
| --- | --- | --- |
| Area# | CommunityArea | Hometo |
| **8** | NearNorthArea | WrigleyField,LincolnParkZoo,LincolnPark, Chicago History Museum,Boystown, North Avenue Beach, etc. |
| **28** | NearWestSide | Garfield Park Conservatory, UnitedCentre, National Museum of MexicanArt, Wicker Park, etc. |
| **32** | TheLoop | The Art Institute of Chicago,CloudGate, Willis Tower, Grant Park, Chi­cagoCulturalCentre,ChicagoTheatre,etc. |
| **33** | NearSouthSide | Adler Planetarium, SheddAquarium,FieldMuseum,NortherlyIsland,Gless-ner House, etc. |
| **76** | O’Hare | O’Hare International Airport, RotundaTower Garden, Schiller Woods, Sky-deck Chicago, Museum of Contempo­rary Art, etc. |

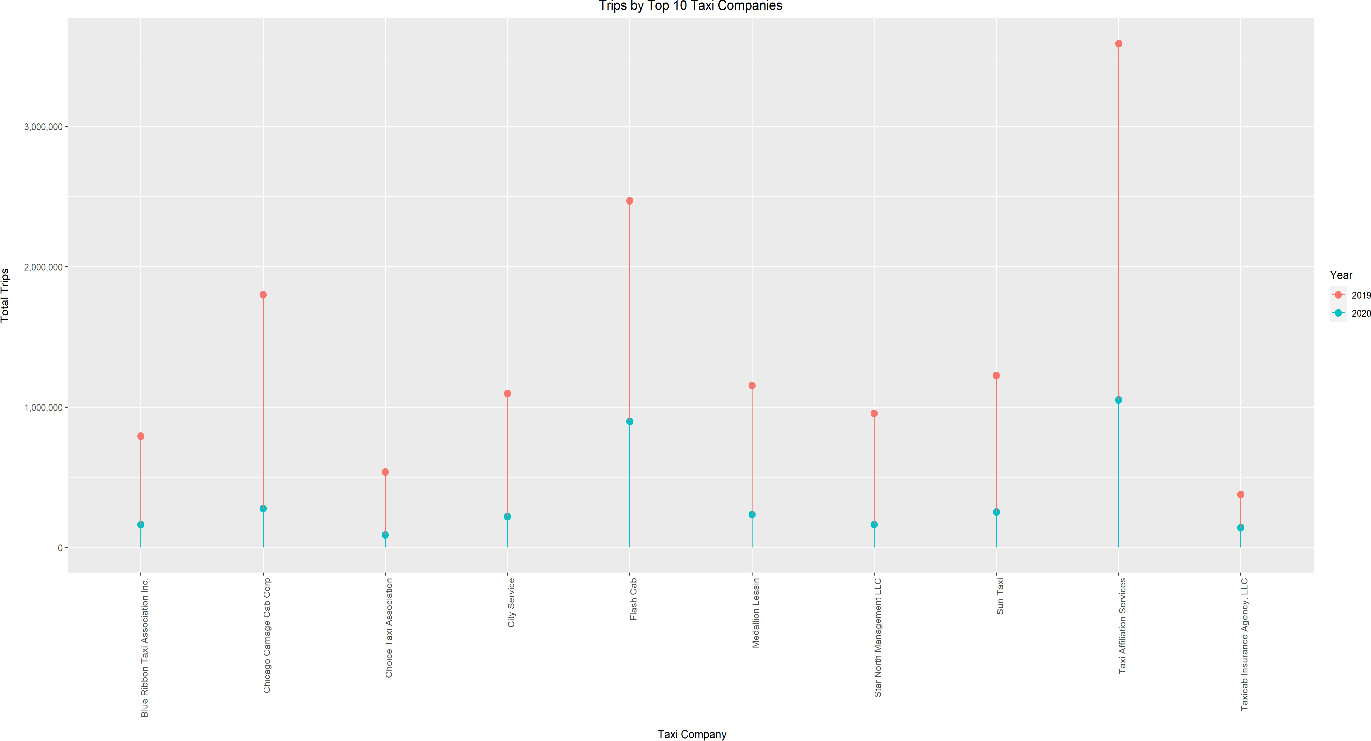
##### TaxiTripsofferedbyTaxiCompanies

ThedatacontainsdetailsofthecompaniesofthevarioustaxitripsmadeinChicago.



Fromtheabovegraph,itcanbeobservedthatthereweremorethan50taxiservice providers in the city during these 2 years.

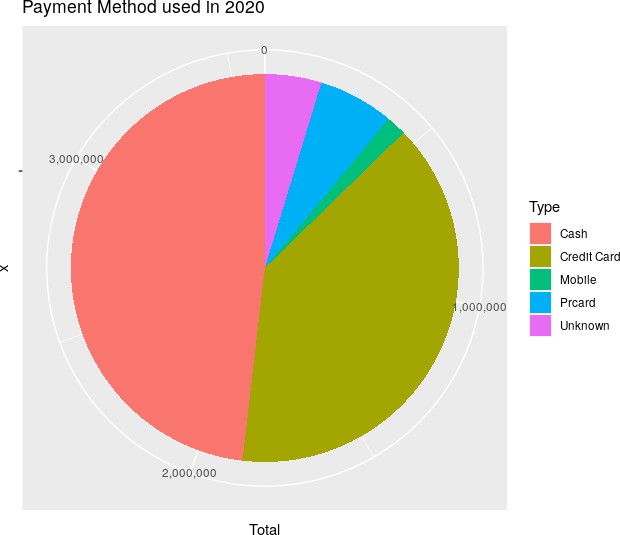
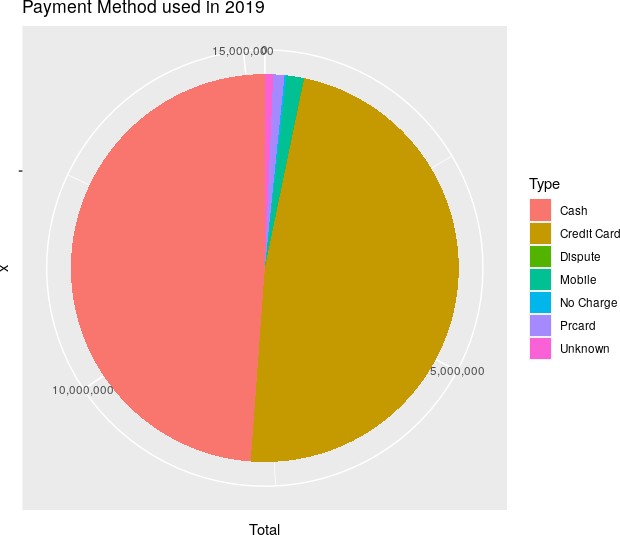
Also,itcanbeobservedthatafewtaxicompanieshaveveryminusculeornotaxi trips in 2019 and a few companies who did not have a taxi trip in2019have managed to offer services post-pandemic.



Though the number of taxi trips reduced considerably, the following taxi compa­nies consistently offered the highest number of rides before and after the pan­demic.

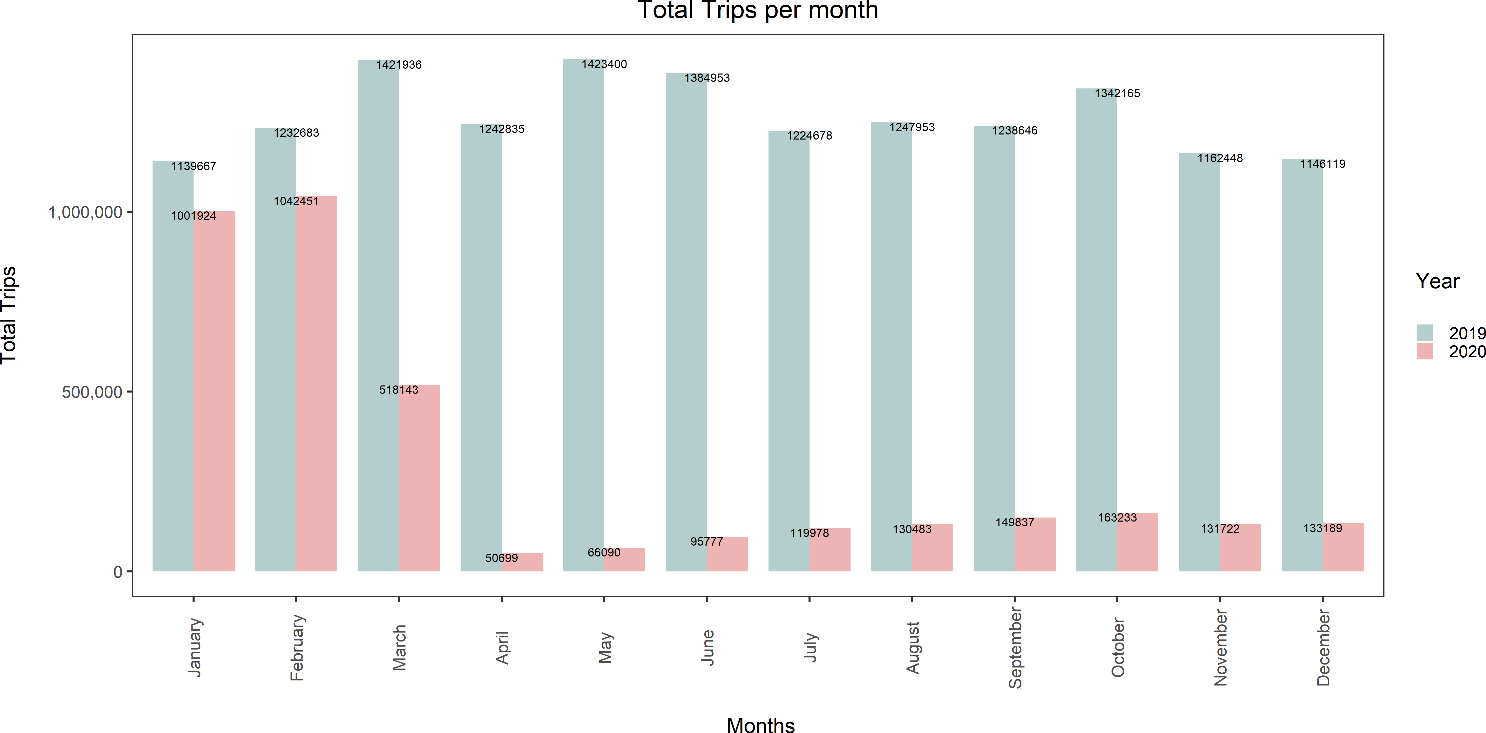
* TaxiAffiliationServices,LLC
* FlashCab
* ChicagoCarriageCabCorporation

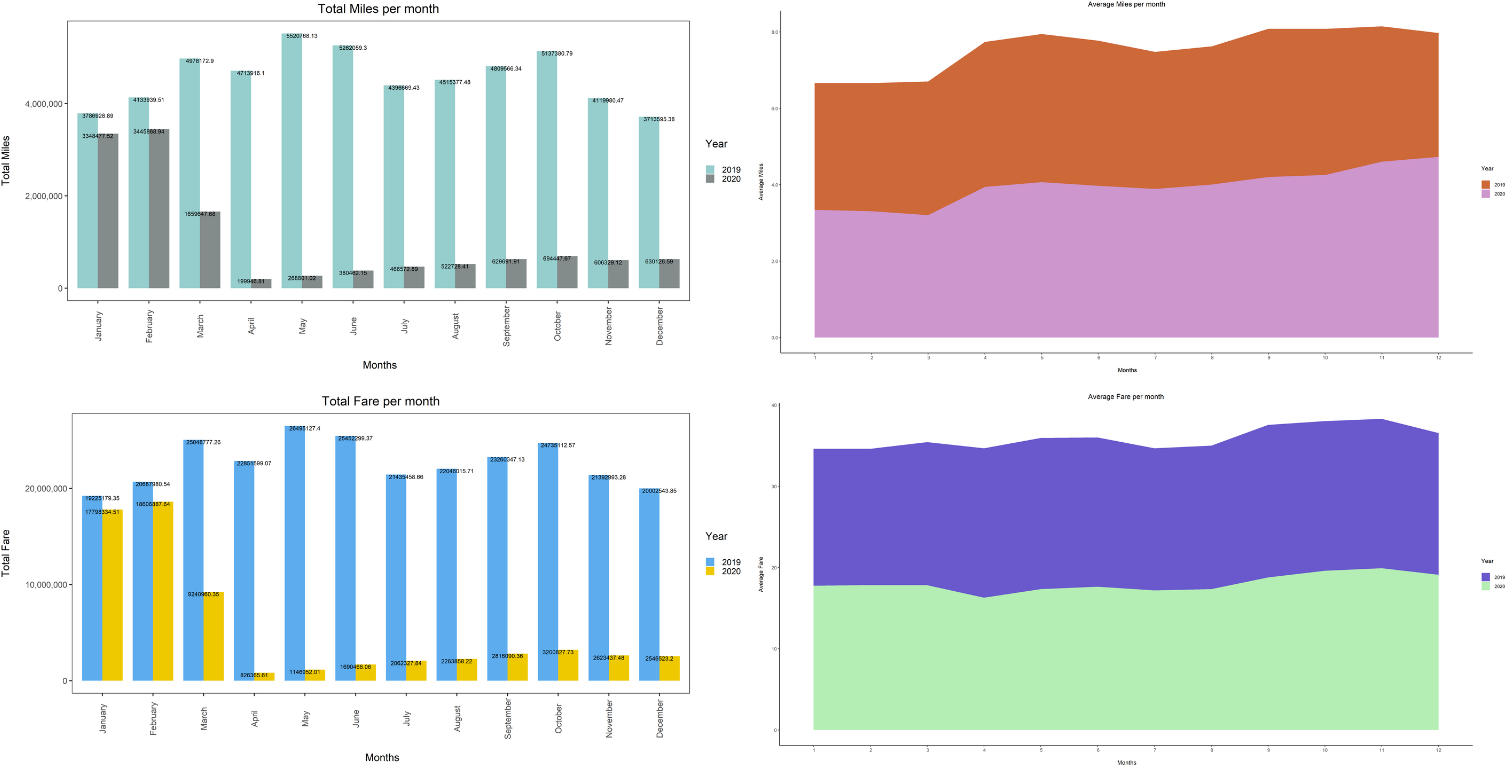
##### ModeofPayment(2019vs2020)

****

Whilecashandcreditcardshavebeenthepreferredmethodsofpaymentinboththe years 2019 and 2020, there is a slight dip in the number of credit card pay­ments after the pandemic. It looks like digital apps have started gaining momen­tum post-pandemic.

##### Numberoftaxitrips,totalmilesandtotalfareacrossmonths(2019vs2020)

****

****

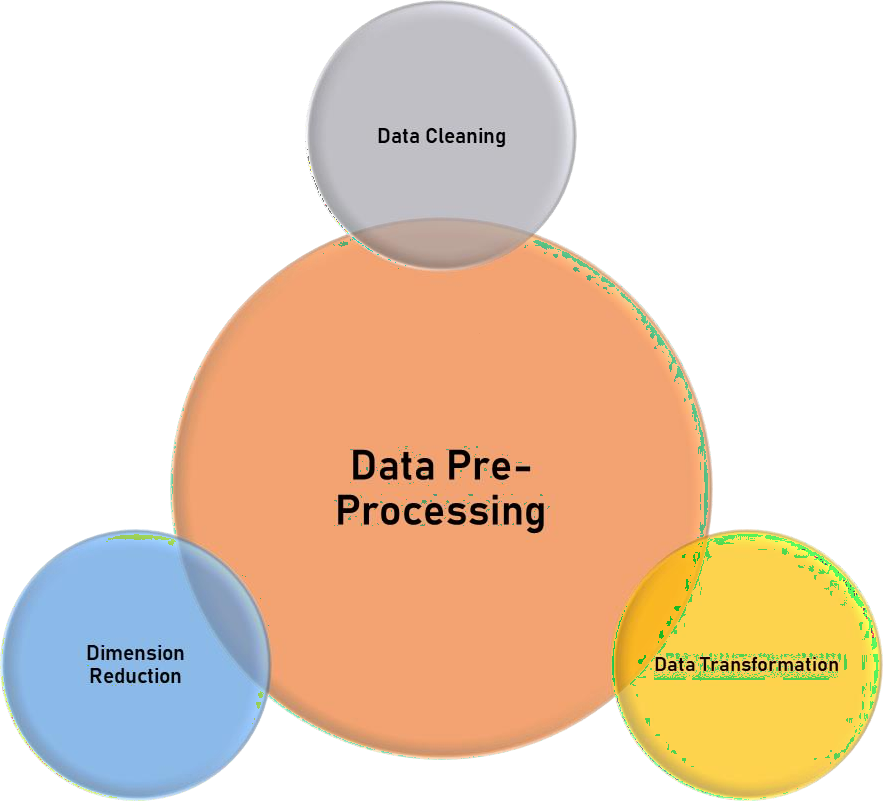
Intheyear2019,thetotalnumberoftripstakeneverymonthremainsstable withMarch,May andJune being thechart-toppers. Onthe other hand, the year2020showsadownwardtrendduetotheoutbreakofCOVID-19asithitsrock-bottominAprilandplateausthroughtherestoftheyear.Theplotsfortotalfareandtotal

miles follow the same trend as the number of taxi trips is directly proportional tothe distance covered on-road and revenue of the taxi industry.

Aninterestingpointtonotehereiseventhoughthebusinessforthetaxiindustryhas been dull in 2020, the mean number of trips taken and the mean trip costcontinuetoremainstable.Ondiggingdeepintothis,wecaninferthatthetaxifarehas either not been increased or the hike is insignificant in the year 2020.

# DataPre-processing

Raw data from an entity isn’t very useful as it is often ambiguous. Data Pre-Pro-cessing is the step wherein data is encoded, manipulated, or dropped so that itmatchestherequirementsofamachinelearningmodel.Owingtothelimitedpro­cessingspeedofourlaptopsandtheenormoussizeofthedataset(around8GB),data is divided into batches and processed separately. Various Data Pre-Pro-cessing measures were implemented to ensure that any irrelevant or discrepantattributes don’t disturb the functioning of the models.



##### DataCleaning

TheChicagotaxitripdatasetconsistsofafairlyhighnumberofrecordswithdatathatisincompleteormissing.Thesevaluesarenotsuitabletothemodelthathasbeen employed, and hence havetoberemoved. We usedR programmingto re­move records containing null and missing values.

##### DimensionReduction

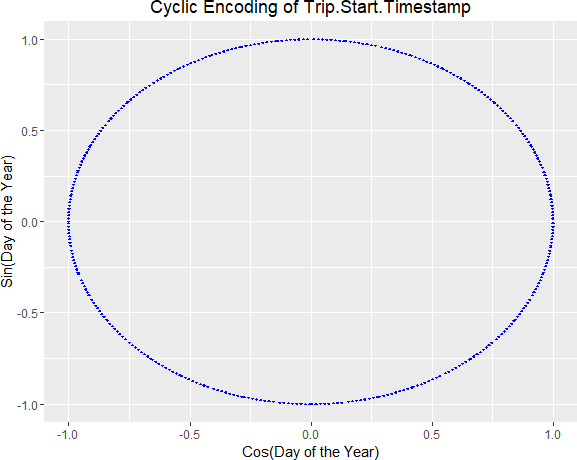
The higher the dimensionality of a dataset, the harder it is to fit it into a model.Atotalof23attributesformsthisdataset,henceitisessentialtoomittheredundantfieldsasitreducesthesizeofthedataset.Weremovedattributesthatdon’tserveany functionality toour business objective. Reducing the dimensionality enabledusbettercomputationaltimetoprocesshighdata.Thefollowingaresomeattrib­utes that were removed from the chosen dataset.

* PickupCensusTract
* DropoffCensusTract
* PickupCentroidLatitude
* PickupCentroidLongitude
* PickupCentroidLocation
* DropoffCentroidLatitude
* DropoffCentroidLongitude
* DropoffCentroidLocation

##### DataCleaning

Input attributes to the model in the relatively clean data set have to be modifiedsothatthemodelunderstandsitanddiscoverstheunderlyingpatternstoproduceuseful insights. The presence of time variables in the taxi dataset calls for cyclicencoding.Toachievethis,thestartdateisfirstextractedfromthestarttimestampof the type character.

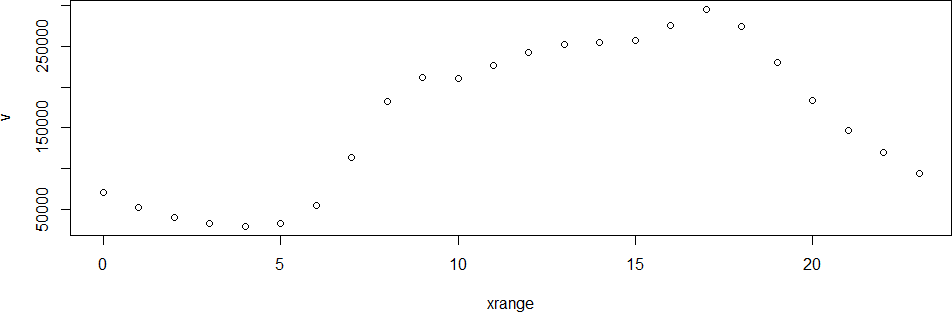
A function is called to numberthe days of theyearfrom 1to 365 (or 366, incaseofaleapyear).Thesineandcosinevaluescorrespondingtoeverydayoftheyearare computed using the functions sin () and cos (). A plot against the sine andcosine counterparts of a day of the year looks like the graph below:



An ideal way to encode a time series object is by using the “cyclic\_encoding()”functionfromapackagecalled“lubridate”.Giventheenormoussizeofthedatasetandtheincorrect datatypeof timevariables, thesaid object didn’t provetobeanoptimal method.

A new set of attributes are introduced to the dataset as a result of manipulatingsome existing attributes, namely – TimeBin, Avg\_Miles and Avg\_Fare.

* TimeBin - It categorizes every record based on the start time. We plottedthe start time for all the trips and divided the day into 5 parts.



|  |  |  |
| --- | --- | --- |
| i.Bin‘1’–12:00AM | to | 4:59AM |
| ii.Bin‘2’– 5:00AM | to | 9:59AM |
| iii.Bin‘3’–10:00AM | to | 2:59PM |
| iv.Bin‘4’– 3:00PM | to | 5:59PM |

v.Bin‘5’– 6:00PM to11:59PM

This classification is donetaking into consideration the hour at which atripbegins. This value is obtained by using an in-built function called “strp-time()”.

* Total\_Trips – This field represents the total number of taxi trips in a givencommunity area on a given date during a specific time interval.
* Total\_Miles-Thisfieldrepresentsthesummationofthedistance(inmiles)coveredduringalltaxitripsinagivencommunityareaonagivendatedur­ing a specific time interval.
* Total\_Fare-Thisfieldrepresentsthesummationofthecostofeverytrip(inUSDollars)foralltaxitripsinagivencommunityareaonagivendateduringa specific time interval.
* Avg\_Miles-Thisfieldrepresentsthemeandistance(inmiles)covereddur­ingalltaxitripsinagivencommunityareaonagivendateduringaspecifictime interval.
* Avg\_Fare-Thisfieldrepresentsthemeancostofeverytrip(inUSDollars)foralltaxitripsinagivencommunityareaonagivendateduringaspecifictime interval.

# Modeling

Oneofthebiggestchallengesthetraditionaltaxicompanyfacesisimprovingtheefficiency of its taxi fleet. A predictive model capable of predicting taxi demandwill help the companies plan ahead of the schedule of their taxis which will helpin reducing wait time for the passengers, improving utilization of the resourcesand also help in predicting traffic “hot-spots”.

For this project, we decided to apply and compare three modeling algorithms,namely:VectorAutoRegression,Multi-LayerPerceptronandRecurrentNeuralNetworks. These algorithms have beenextensively studiedand applied to pre­dict time series data.

Wealsodecidedtocreateseparatemodelsforeachcommunityareasinceitcan beinferred that eachcommunity area hasits time series andinherit pat­terns. In the future, data features extracted from other community areas canalso be used to better model the taxi demand prediction.

#### VectorAutoRegression(VAR)

VARmodelshelpcapturerelationshipsbetweendifferentfeaturesinatimeseriesdataset. It uses the past values of the variable, called lag, past values of otherdependent variables and error terms to predict the next variables in the series.

VARmodelshavebeenusedinnaturalsciences,financesectorsandeconomiesas they mostly deal with time-series data. Other areas of applications include:

* 1. Datadescription
  2. Forecasting
  3. Structuralinference
  4. Policyanalysis

##### AdvantagesofVAR

1. Forecastingarelatedvariablecollectionwherenoextrainterpretationis required.
2. Testingisdoneifonlyonevariableisusefulforforecasting.

#### DeepNeuralNetworks-MultilayerPerceptron

Adeepneuralnetworkisatypeofartificialneuralnetworkthathasmultiplehiddenlayersofunitsbetweenitsinputandoutputlayers.DNNs,likeshallowANNs,canmodel complex non-linear relationships. DNN architectures produce composi­tional models in which the object is represented as a layered composition of im­age primitives. The additional layers allow for the composition of features fromlowerlayers,allowingforthemodelingofcomplexdatainfewerunitsthanasim­ilarly performing shallow network.

Neuralnetworksweretheprimechoiceofmodelingalgorithmsforthisprojectduetotheybeinguniversalapproximators,especiallyfornon-linearrelationships.Dueto their nature, neural networks are specially used to model time series data.

##### Architecture

ThedevelopmentofMLPnetworkshas2mainagendas:Architecture&Training.The architecture of MLP is very important because a smaller number of connec­tions may not be able to capture all the relationships within the data features,whereas more connections may cause an over-fitting of the training data.

The multilayer perceptron is one of the most prevalent types of deep neural net­works.Deepfeedforwardnetworksormultilayerperceptronsarethefoundationaldeep learning models.

Thesemodelsarenamedfeedforwardbecauseinformationflowsviathefunctionbeing evaluatedfrom x, throughthe intermediatecomputations necessary tode­finef,andfinally theoutputy.Therearenofeedbacklinks, thereforethemodel’soutputs are not transmitted back into it.

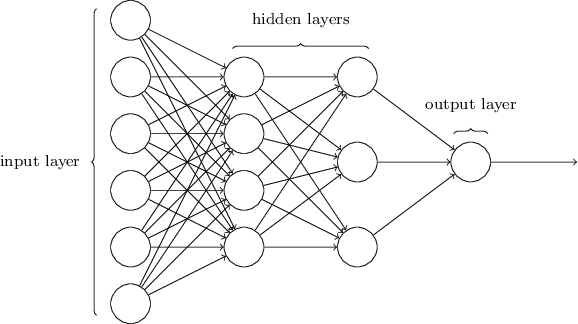


Fig1:MultilayerPerceptronArchitecture**Disadvantages of MLP**

1. Computationsareverytime-consumingandsometimesbecomecompli­cated because of complex connections.
2. Neuralnetworksareblackboxes;It'shardtointerprettheirlearnednetworkand how the training algorithm arrived at that particular structure.

#### RecurrentNeuralNetwork(RNN)

Recurrentneuralnetworksareaclassofneuralnetworksthatallowpreviousout­puts to be used as inputs while having hidden layers. RNNs models are exten­sively used in natural language processing, speech recognition, sequential pre­diction, etc.

Recurrent neural networks follow backpropagation through the time algorithmtodeterminethegradients.Theprinciplesofbackpropagationtimearethesameasthegeneralbackpropagation.Themodeltrainsitselfbycalculatingerrorsfromitsoutput layers and shifts back to its input layer.

##### TypesofRNN

1. One-to-One
2. One-to-Many
3. Many-to-One
4. Many-to-Many

##### feed forward neural networkWorkingofRNN

Fig2:RNN

In a feed-forward network, when the training data is given to the model it goesthrough the hidden layers evaluating through the activation layers and displaysthe output through the output layers. This network doesn’t touch the nodes onceagain after going through it.

Feed forward networks are not efficient in predicting the new values becauseitconsiders only the input which comes from train data.

RNN completes its information in a cycle through a loop. When producing theoutput,RNNlearnsfromitscurrentinputandconsiderstheinputvaluewhichwasgiven previously.

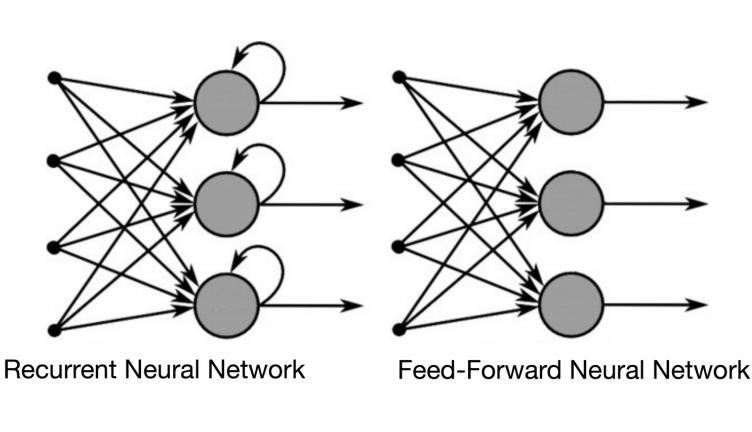


Fig:ComparisonbetweenRNN&Feedforwardnetwork

##### AdvantagesofRNN

1. Canprocesstheinputofanylength.
2. Modelsizeremainsconstant.
3. Weightsaresharedaccordingtotime.

##### DisadvantagesofRNN

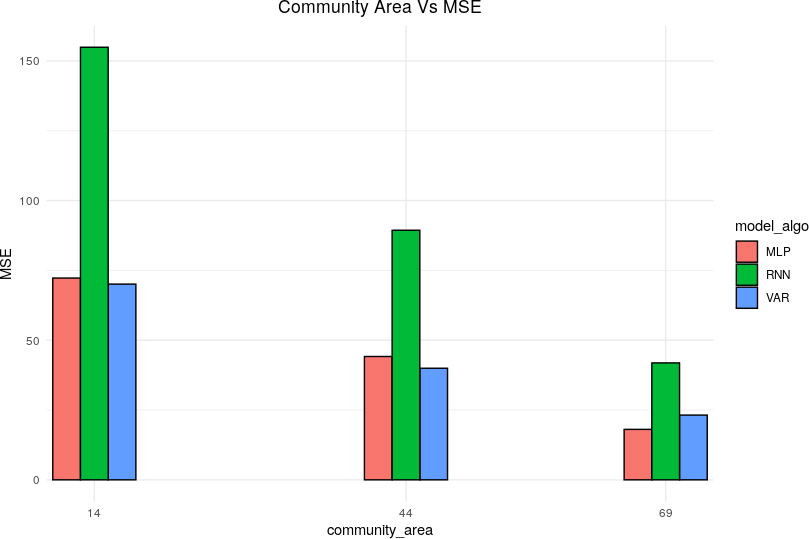
1. TrainingRNN’scanbeverycomputationallyexpensive.
2. Cannotconsideranyfutureinputforthecurrentstate.

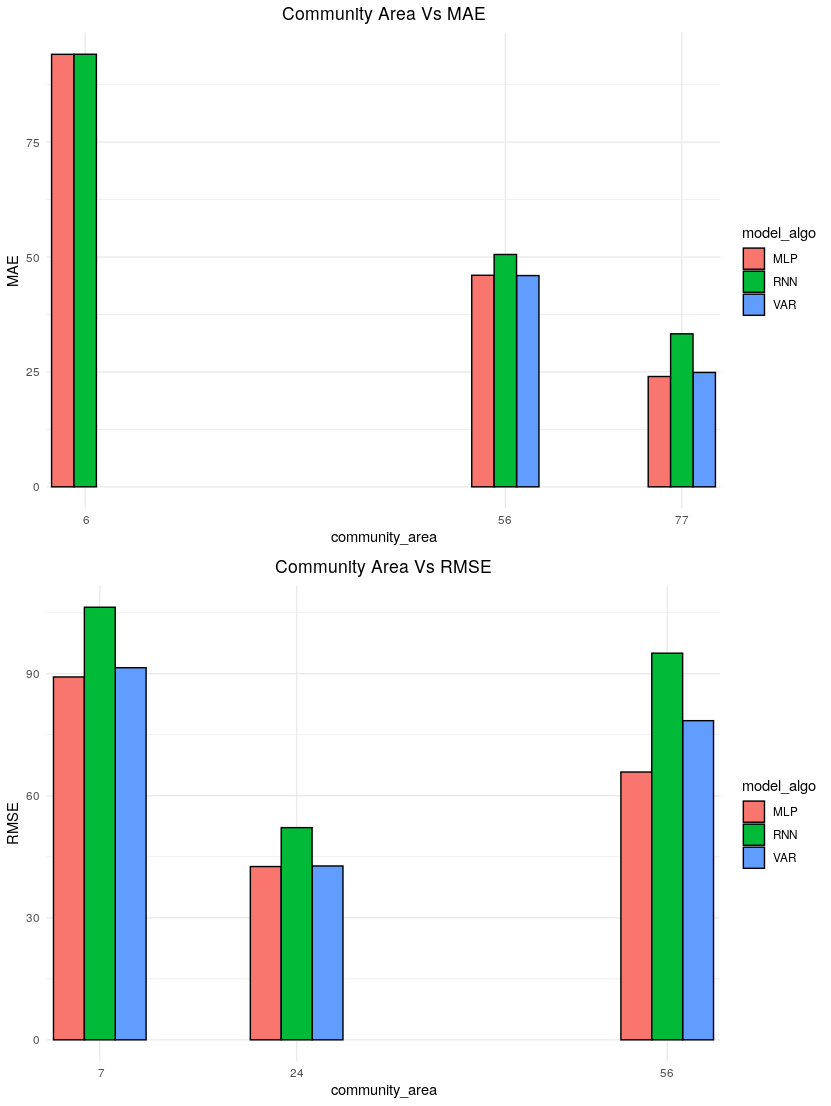
**MLPvsRNNvsVAR**

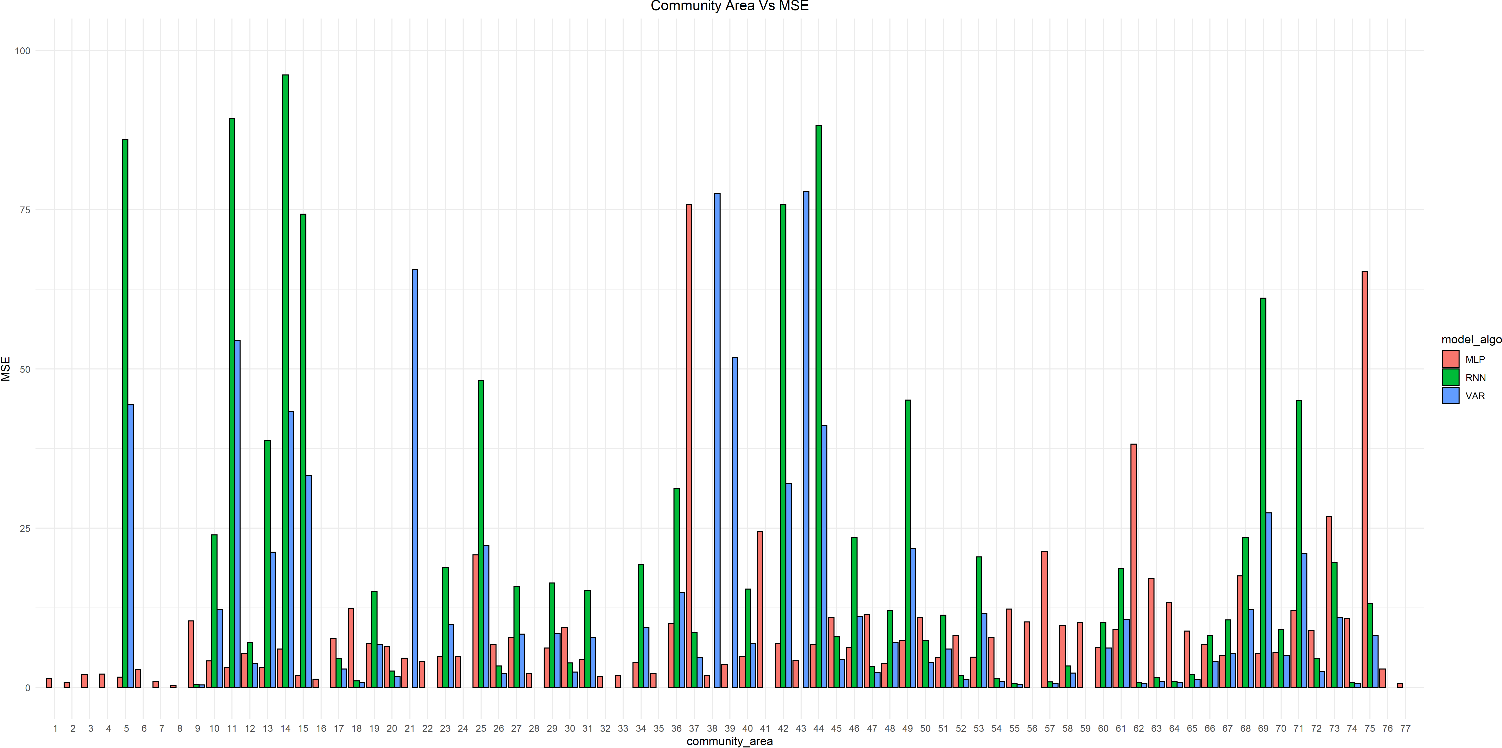
|  |  |  |
| --- | --- | --- |
| **Multi-LayerPerceptron** | **RecurrentNeuralNetwork** | **VectorAutoRegres­sions** |
| MLP is the foundation ofmore complex deep neu­ral networks | RNNisdesignedforse­quence-related problems. | VARs are generallyusedtomodellinearre­lationships betweenvariables over time. |
| Can model non-linear re­lationships | Can model non-linear re­lationships as well as cy­clic patterns | Notusefulformodelingnon-linear relation­ships |
| Doesnottakepreviousvalues into account | Runs a feedback loop, asa result, previous valuecan influence present val­ues | Uses lag variablestoaccount for cyclic pat­terns. |

# Evaluation

We have calculated Mean Absolute Error, Mean Squared Error and Root MeanSquaredErrorfor3algorithmsforall77communityareas.Also,foreachevalu­ation metric, we are comparing the modeling algorithms for only the top 3 com­munity areas which show the greatest variance, to better visualize the model’scomparative performances.







The graphs clearly show that the metrics are generally on the higher side forRNNs, followed by VAR and MLP. Interestingly, we expected Recurrent NeuralNetworks to perform better than the other two algorithms, however, the metricsshowadifferentstory.However,thisshouldn’tbededucedtothefactthatRNNsare a weaker selection. Complex neural networks like those require a preciseapproach to data preparation and hyperparameter selection, success in achiev­ing them will result in a significant accuracy improvement.

MLP, being a simple deep neural network, performed better than RNN, whichrequiresamorepreciseapproachandVAR,sinceVARcanonlyrecognizelin­ear relationships.

# Reflection

One of the biggest challenges we faced as a team was working with a large da­taset. Our approach had to be modified in the later stages of the project duetoourtaskofhandlinglargedatasets.WemadeuseofalibraryinRnamed“data.ta-ble” which aids in fast processing of large data, limited to available RAM.

WealsoobservedthatusingcomplexneuralnetworkslikeRNNsrequiresamuchcareful approach to data preparation, feature selection and hyperparameter se­lection. In a practical business scenario, a simple deep neural network suchasMLP can satisfy business requirements whilst reasonably compromising on pre­diction accuracy.

# ConclusionandFutureScope

WeperformedastatisticalprojectontheChicagotaxidatasetusingCRISP-DM.Through Exploratory Data Analysis, wediscovered andcompared trends in pre-COVID and post-COVID eras. We pre-processed and modeled the dataset topredict taxi demand using 3 modeling algorithms, and compared their perfor­mances.

Lastly, we had experience working on a very large dataset and adapted ourcodetoworkonthem.Eventhoughthedatasetweworkedonwasfairlylarge(around 8 GB), it was just a small sample of the overall dataset, starting from2016.

In future projects, we would like to further refine our approach towards modelingRecurrentNeuralNetworks,tofullyutilizetheirpotential.Wecouldalsocombinethe dataset with datasets to capture external signals which might be affectingthe demand such as weather, events and festivals, traffic data, etc.

### References

1. Lupkin, J. M. (n.d.). Encyclopedia of Chicago. Retrieved from<http://www.encyclopedia.chicagohistory.org/pages/1232.html>
2. Tribune, C. (n.d.). 1920—Taxi Wars. Chicago: Chicago Tribune. Retrievedfrom https://chicagology.com/notorious-chicago/1920taxiwars/
3. Uber. (n.d.). UberDATA: Uber’s Economic Impact on the City of Chicago.Retrieved from https://[www.uber.com/blog/chicago/uberdata-ubers-](http://www.uber.com/blog/chicago/uberdata-ubers-)economic-impact-on-the-city-of-chicago/