Mining Uber Data for Added Services

Mahmoud Helmy
mhelm081@uottawa.ca

Minah Ghanem
mghan079@uOttawa.ca

Mohammed Elnamory
melna086@uottawa.cal

Sarah Elmasry
selma083@uottawa.ca

***Abstract*—In this paper we are going to explore multiple added services to Uber trips and how its services can be improved using data mining.**

For the first problem statement, we presented a study case that is concerned with predicting Uber's mean travel time. The study is comprehensive and includes all the stages from data collecting to evaluating and coming up with conclusions. At the end of the first section we represented our criticism and our contribution.

As for the second problem statement, we are going to predict Uber's popular locations and analyze the public transportation network in different areas using real-time data analytics on the Kubernetes environment. We criticize the current solution, and propose our framework which is based on a different clustering algorithm and the use of AWS services.

We can see that a lot of studies are based on improving Uber services using its enormous data. These studies aim at analyzing the usage patterns to determine where Uber should offer or focus its Service in order to increase their profits from real-time analysis. Nevertheless, less studies were focused on the safety of the trip. Actually we found a huge gap and less information in safety services so at the end of this paper we proposed another framework solution that focuses on the safety of Uber trips.

***Index Terms*—business data processing, computer vision, clustering methods, image analysis, artificial intelligence**

I. Introduction

If we take a look at the last 5 years, we will see how Uber's annual revenue increases year by year. A simple app with Easy user interface, you just press a button to book an Uber, select the pickup location, request a car, choose your payment method and go for your ride. This straightforward process is not the same in the back-stage, a lot is going on behind the scene. "Uber lives or dies by data. "- said Spencer, a former Uber driver. Uber transports people around the world without owning any cars, Uber owns data.

In 2020, Uber had 93 million active users in +10,000 cities across 71 countries which makes Uber the fastest growing startup standing at the top. and Here we can imagine the huge raw data that Uber has and how they can use this data to predict,find insights and study user behaviour.

II. Problem Formulation

- The path between the pickup point and the destination is not the only factor to predict the travel time. Weather, traffic congestion and subway density are big factors who affects on our daily travel time, Even the weekend and holidays are important. To predict the trip time we should take in our considerations introgenous and exogenous variables.
- Uber is using real-time Big Data to perfect its processes: 1-from calculating Uber's pricing to finding the optimal positioning of taxis to maximize profits.
  2-Real-time data analysis is very challenging for the implementation . Implementation of real-time data analysis by Uber to identify their popular pickups would be advantageous. It will require high-performance platform to run their application.
  As the increase of Uber's users and driver although the size of the company this will demand more computional power and more father ways and also reducing the cost too.
- We hear Uber riders and drivers share fears about safety during their trip. We always hear about murders, kidnaps, and sexual assault in news. Also, many incidents happen every day and lots of death.

In the following sections, we will discuss those two main problem. Also,our point of view of how research in and what do we think is missing. this area is moving

III. Travel Time Prediction [1]

A. Summary

Our problem statement here is to predict the mean travel time in Uber trips.This study case makes use of multiple sources of data. The study's main source of data is "Uber Movement" which is a website that employs Uber's riding data. All the trips' origin was set at Washington D.C. city center and the destination locations include holistic areas (zones) also in DC. Other sources of data are also included like the daily weather conditions so we can analyze how it affects the traffic.

As for the data processing and preparation phase, multiple features were derived to enhance the study. For example the shortest path is calculated between the two nodes. Having only the geometry polygons of origins and destinations from the Uber Movement website, we used the Betweenness centrality algorithm to estimate the points of both destination and source points. Then by using the "Networkx" library in the Python environment we calculate the shortest path. Another feature to mention is the density of the street, and it is calculated using "OpenStreetMap (OSM)" Library.

Descriptive analytics also was represented in this study case, to explore the relations between the current features and the mean travel time of the trip. For example, the correlation between the mean travel time and the shortest path was explored and the results came in as a corn distribution, which means that mean travel time intends to increase when the travel distances become longer. The correlation between the street densities, working day VS vacation day, temperature and the mean travel time were all explored in this study case.

As for the modeling part, two models were developed. One for the traveling behaviours in the working days and the other one is for the non-working days. Before the modeling phase, a Sequential Feature Selection algorithm was used to enhance the model. It works by selecting a subset from the features that is most relevant to the model.Then two algorithms were used: the Huber Robust and the Random Forest. They were both tuned by using a grid search algorithm to try and match different combinations of hyper parameters. At the end, the study showed that the Random forest model in both the working and non-working models scored better than the Huber Robust model with a prediction error of (MSE:-0.0081) compared to a (MSE:-0.0909) of the Huber Robust model.

B. Criticism

- In the proposed paper, he used the origin of all Uber trips was set at Washington D.C. city center and Destination locations include holistic areas (zones) that Uber covers in D.C. which makes our area predict very limit, we should always start with the same origin to the selected area.
- According to NetworkX.shortest_path() [2] who calculate the shortest path between two points.

  The paper does not take in advance the weight of each node. For example, the shortest path out for the previous maybe has a small road and many traffic lights, the result the trip takes more time than predict.

  The relation between the trip time and street density is direct proportion.

  When added weight for each node with the density respective.The output path will be more accurate.
- Paper's data covers all weather and Uber data for January, February and March 2018. This dataset is facing many problem.

  First, the weather for those three months is probably the same, The model has not the chance to train on another kind of four seasons. Moreover, it does not test in the summer season when the schools are on vacation so the most probability the time travel will have another curve. It will have a huge change if the area has many schools for example.

  For conclusion, the model has many weak points in the time series.
- The paper chose to try Sequential Forward Selection to select a subset of features with The Huber Robust Regression without explanation. why if he used Stepwise Regression or Backward Elimination for better results.
- According to the proposed paper, he has tried random forest and The Huber Robust Regression.

  Another Comparison with the Support Vector Machine Algorithm is an impressive addition. On the other hand, The Deep Learning Model with those large number of parameter it may be better.
- The criteria of choosing a holiday depend on the USA calendar which makes the model limited. For instance, The Middle East holidays are different and do not exist in the model's parameters.

C. Contribution

In our contribution, we used the Uber Movement website to get the data of a different city "Madrid" and used the uber riding data for three months (October-November -December). The source of all the trips was Código Postal 28001 and the destination was Delicias, Madrid.

We used this data to perform some analytical tasks on it and to discover the correlation between the target and predictive variables. In figure 12, we explore the effect of the day of the week on the mean travel time. We can see that Saturday-Sunday-Monday have the least mean traveling time, which concludes that the average traveling time tends to decrease when it is a non-working day. And the mean Travel time is at its peak in the middle of the week where most business activities are done.

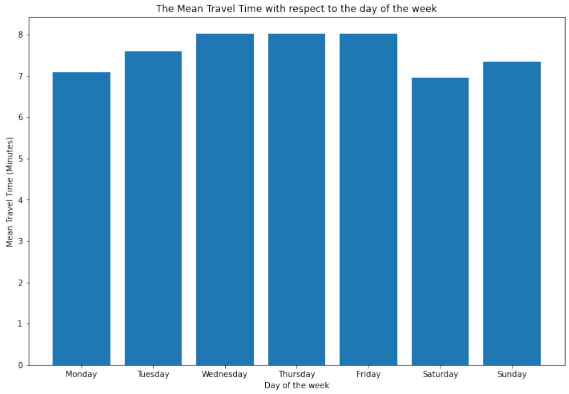
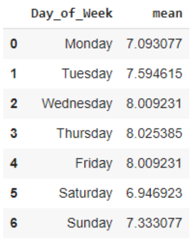


Fig. 1. Mean Travel Time plot with respect to week day



	Day_of_Week	mean
0	Monday	7.093077
1	Tuesday	7.594615
2	Wednesday	8.009231
3	Thursday	8.025385
4	Friday	8.009231
5	Saturday	6.946923
6	Sunday	7.333077

Fig. 2. Mean Travel Time values with respect to week day

In Figure 3, We have also explored the effect of each month on the mean travel time and found that December has the most diverse distribution compared to the other two months as it has lots of holidays.

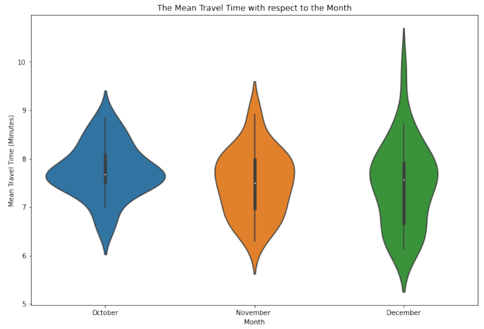


Fig. 3. Mean Travel Time values with respect to Month

IV. Real time Data Analysis [3]

A. Summary

There is a growing demand for Big Data applications to extract and evaluate information, which will provide the necessary knowledge that will help us make important rational decisions. Big Data analytics is the method of analyzing massive data sets to highlight trends and patterns.

B. Criticism

- The choice of algorithm:
  He mentioned in the implementation, that he applied K-means clustering algorithm on the dataset, without stating why he chose this model over the others,despite that K-Means has many Drawbacks like :
  - the number of clusters must be decided before the analysis. This involves a combination of common sense,domain knowledge, and statistical tools.
  - calculating a the best k for a large dataset could potentially crash a computer due to the computational load and the limits of RAM.
  - Randomly selection different places from which to develop clusters. This can be good or bad depending on where the algorithm chooses to begin at.
  - The order of the data has an impact on the final results.
  - It is good in capturing structure of the data if clusters have a spherical-like shape. It always try to construct a nice spherical shape around the centroid. if the clusters have a complicated geometric shapes, kmeans does a poor job in clustering the data. [4]

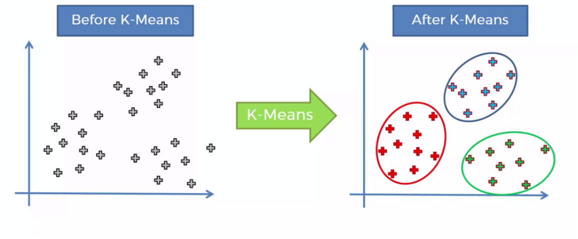


Fig. 4. k-Means Algorithm

  - The algorithm doesn't let data points that are far-away from each other share the same cluster even though they obviously belong to the same cluster. [5]

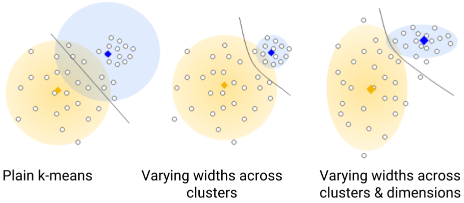

Plain k-means
Varying widths across clusters
Varying widths across clusters & dimensions

Fig. 5. k-Means Algorithm Clustering

- The choice of Cloud providers: It is stated on the google cloud website that Dataproc [6] is a fully managed and highly scalable service for running Apache Spark, Apache Flink, Presto, and 30+ open source tools and frameworks. Use Dataproc for data lake modernization, ETL, and secure data science, at planet scale, fully integrated with Google Cloud, at a fraction of the cost, and he didn't mention why:
  - He didn't compare or mention his experiment with other cloud providers like azure or AWS .
  - He didn't explain why he chose google cloud platform over all other providers .
  - He mentioned only that google cloud platform is super fast and easy use platform without giving any statistics or provide an explanation for his point of view ,which was very weak argument.

Fig. 6. Giant Cloud providers in the industry

C. Proposed Solutions:

- From the above mentioned drawbacks of kmeans , we propose the use of other clustering algorithm for example DBSCAN .
  in comparison with K-means doesn't need to define the number of clusters at the beginning also DBSCAN has a notion of noise, and is robust to outliers. [7]

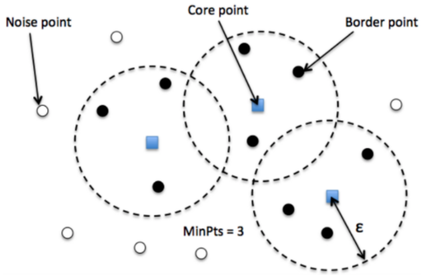


Fig. 7. DBSCAN Algorithm

- Nowadays the choose of a cloud provider over the others is a very hard task.As you can see in the picture that there are many cloud providers and the competition is at a very high level:

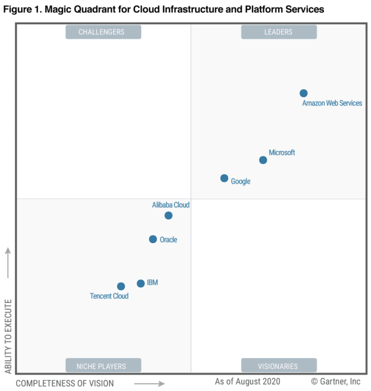


Fig. 8. Cloud providers in the market

  - Amazon, Microsoft and Google dominate the public cloud landscape providing the safest, flexible and reliable cloud services. Their respective cloud platforms, AWS, Azure and GCP offer clients a range of storage, computing and networking options.
  - Some of the features common among the three platforms include instant provisioning, self-service, autoscaling, identity management, security and compliance, among others.
  - At present, AWS can be considered to be much bigger than both Azure and GCP in terms of functionality and maturity [8]. So he should at least compare AWS , azure and google cloud for the same service , so we can conclude that the champion cloud platform should be applied with this specific task later.

Our framework is a solution based on AWS services:

- Data ingestion – Enables real-time data ingestion from either an IoT device or data uploaded by the user, and real-time data storage on a data lake. This functionality is specifically tailored for situations where there is a need for storing and organizing large amounts of real-time data on a data lake.

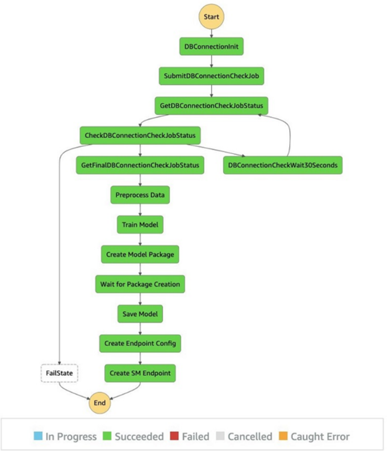


Fig. 9. Framework

- Scheduled model refresh – Provides scheduling and orchestrating ML workflows with data that is stored on a data lake, as well as training and deployment using AutoML capabilities.
- Real-time model inference – Enables getting real-time predictions from the model that is trained and deployed in the previous step.

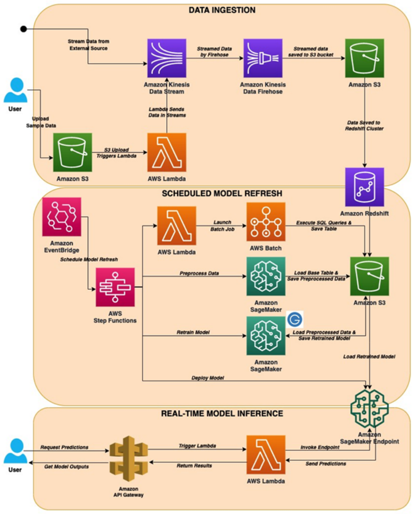


Fig. 10. AWS Pipeline

Scheduled model refresh:

- In this module, you can schedule events using EventBridge, which is a serverless event bus that makes it easy to build event-driven applications. In this solution, we use EventBridge as a scheduler to regularly run the ML pipeline to refresh the model. This keeps the ML model up to date.
- AWS Batch to run queries to Amazon Redshift to ETL data.
- Data preprossessing using Amazon SageMaker.

- Training and deploying ML models using Amazon SageMaker.

The inference module:

- launches a REST API using Amazon API Gateway with Lambda integration, allowing you to immediately get real-time inference on the deployed AutoGluon model.
- The Lambda function accepts user input via the REST API and API Gateway, converts the input, and communicates with the Amazon SageMaker endpoint to obtain predictions from the trained model.

V. Missing points: Uber Safety

According to Research shows that rideshare services account for about a 3 percent increase in traffic-accident fatalities since 2011. That's about 987 deaths per day and due to the increasing number of rideshare vehicles on the road.

Also, Uber vehicles were involved in 97 fatal crashes between 2017 and 2018, leading to 107 deaths. Of that number, 21 percent of the crash victims were the rider, 21 percent were the driver, and the remaining 58 percent were third parties. [9]

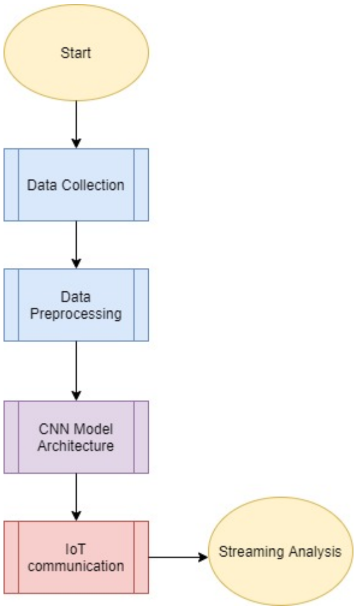


Fig. 11. Safety Uber Framework

The proposed Framework in figure 5 presents the flowchart of our solution.

- Collecting different images of driver and user behaviour, for example, driver is falling asleep,eating or the driver/user is harassing a female.
- Applying data preprocessing like segmentation and image filtering , also, we should classify dataset if it is abnormal behaviour or normal behaviour.
- Applying CNN model to our dataset.
- Communication between the camera device and the application using ioT technology.
- Set a alarm if the model predict abnormal behaviour.
- Send the trip details to police or Uber office to keep tracking on.

VI. Conclusion

In this paper we came across multiple challenges in increasing the added services of Uber.

We discussed the identification problem of Uber popular locations using real-time and suggested another framework to solve the problem based on the drawbacks of the proposed solution.

Another problem we came across is the prediction of the mean time travel in Uber trips and what are the main factors that affect the mean time of travel. We also tried to explore the dataset based on another city (Madrid) than the proposed paper. In the end, we saw that a lot of research is based on increasing the profit of uber services and increasing its quality, however less research is done on the safety of the passengers and how to make more safe trips. So, as an initiative, we decided to suggest a solution to address that perspective of Uber added services.

References

[1] S. Shokoohyar, A. Sobhani, R. Malhotra, and W. Liang, "Travel time prediction in ride-sourcing networks: A case study for machine learning applications," vol. 26, p. 1, 02 2020.

[2] N. Library. (2020) Shortest path algorithm. [Online]. Available: https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algor

[3] T. Gunawardena and K. Jayasena, "Real-time uber data analysis of popular uber locations in kubernetes environment," in *2020 5th International Conference on Information Technology Research (ICITR)*, 2020, pp. 1–6.

[4] I. Dabbura. (2018) K-means clustering: Algorithm, applications, evaluation methods, and drawbacks. [Online]. Available: https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages

[5] G. Developer. (2020) k-means advantages and disadvantages. [Online]. Available: https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages

[6] Google. Dataproc. [Online]. Available: https://cloud.google.com/dataproc

[7] wikipedia. Dbscan. [Online]. Available: https://en.wikipedia.org/wiki/DBSCAN

[8] veritis. Aws vs azure vs gcp – the cloud platform of your choice? [Online]. Available: https://www.veritis.com/blog/aws-vs-azure-vs-gcp-the-cloud-platform-of-your-choice/

[9] brooks law group. (2018) Uber accidents. [Online]. Available: https://www.brookslawgroup.com/car-accident-lawyer/uber-lyft-accidents/uber-and-lyft-crash-stats/