SUMMER INTERNSHIP

ANALYTICS VIDYA
INTRODUCTION TO DATASCIENCE

A Training Report

Submitted in partial fulfillment of the requirements for the award of degree

B-TECH (CSE)

COURSE : CSE443

SPECIALIZATION: DATASCIENCE

SUBMITTED TO:

LOVELY PROFESSIONAL UNIVERSITY

From 05/04/2020 to 06/13/2020

SUBMITTED BY

NAME OF STUDENT : Manikranth Reddy

REG NO : 11809502

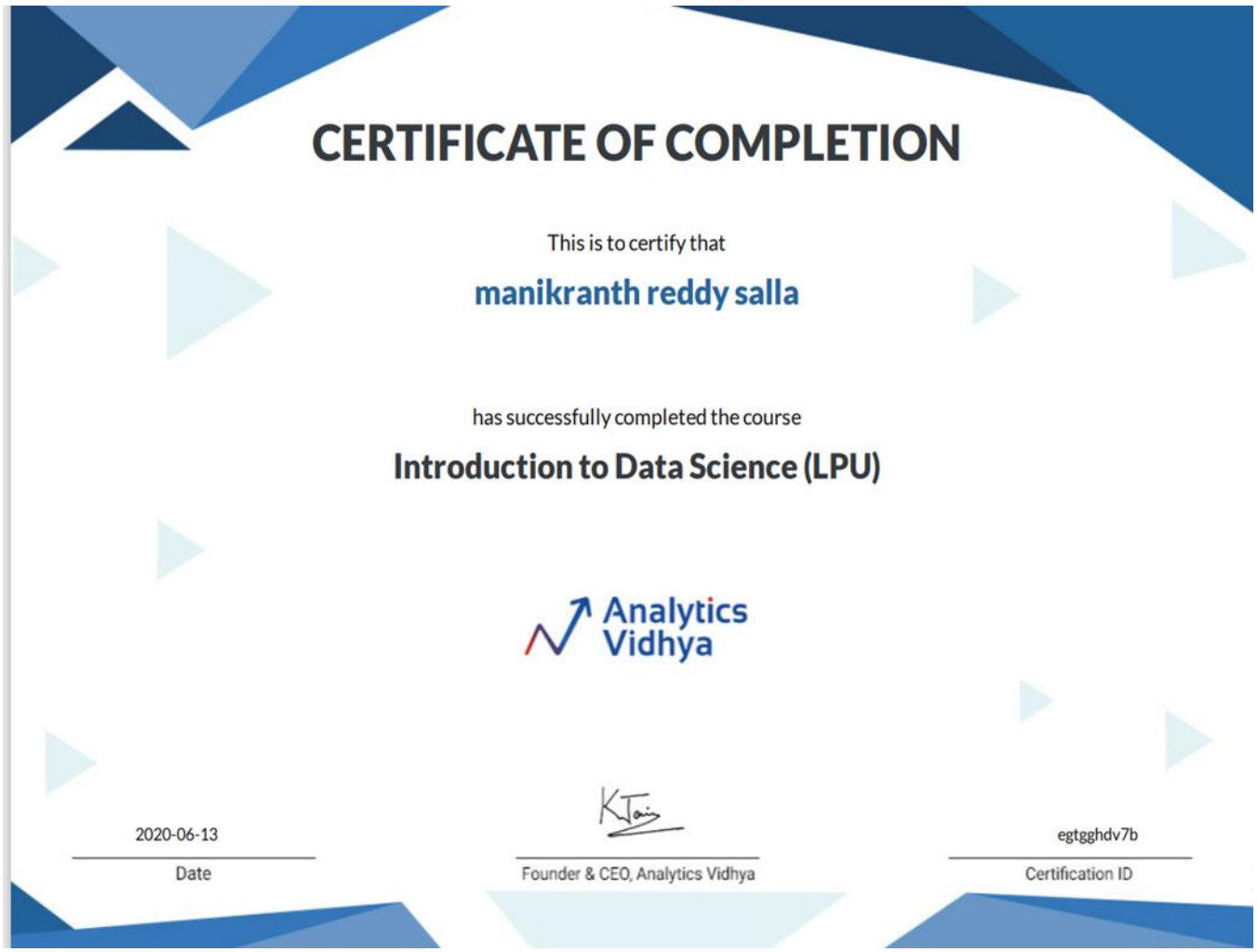
SIGNATURE OF STUDENT : manikranth

I, **Manikranth reddy (11809502)**hereby declare that the work done by me on "**Introduction To Data science**" from **May 2020** to **June 2020**, is a record of original work for the partial fulfillment of the requirements for the award of the degree, **B-tech(CSE).**

Name of the Student (Registration Number) : Manikranthreddy(11809502)

Signature of the student: manikranth

Dated:26 october 2020



INTRODUCTION OF THE PROJECT UNDERTAKEN

I have done my summer internship in **ANALYTICS VIDYA** organization on INTRODUCTION TO DATASCIENCE .This internship consists both training and project work .I have gone through the training proved by the organization . Then by using all the methodologies/process which has introduced in my training I have done my project work of PREDICTION MODEL(NYC) . In training session I have done assignment after every new topic which have given by team.

The assignment have helped to build the models in my project .it made easy to understand new methods and introduced new concept in Data science .By this project I have came to know many new terminologies in Data science like bigdata ,business intelligence ,predictive modeling , managing information system(MIS),machine learning and many more .

I have developed python programming knowledge in this project which has been used in model building and my assignments are related to python libraries .I have learn designing Dataframe ,graphs of different formats like Seaborn ,matplotlib ,standard libraries ,handling ,aggregation ,Subsetting ,sorting data and etc.

This project included new terminology that is Machine learing which is the way for new evolutions .Project mainly includes the terms of machine learning steps to implement the models required for project strategies . It includes statistics and many other procedures to compute the models and designing the project.

Data sciences includes many projects which are result in new technologies with help of machine learning and Artificial Intelligence ,today there is wide change in technology is because of data science . In our day to day life we use data science for predicting the growth of customers, climate records , population prediction , business growth predictions and there are many application of data science.

In no particular order of priority or importance, these are:

Business domain

Mathematics (includes statistics and probability)

Computer science (e.g., software/data architecture and engineering)

Communication (both written and verbal)

My role in this project is to understand and grasp the things to build the allocated model ad to learn the actual uses and application of data sciences and to complete the project by own. I have completed my training which help to understand the things of data science and in project models.
Summarize data.
Understand and use some statistics and mathematical technique.
Prepare data visualizations and reports.

INTRODUCTION OF THE ORGANISATION

Analytics Vidhya is on a mission to create the next generation data science ecosystem. It aim to make data science knowledge accessible to as many people as possible , play an active role in enabling people to create products enabled by Data Science and Machine Learning.

Visibility in front of our community—They have an awesome community passionate about data science and machine learning. Millions of people across the globe use Analytics Vidhya as their source of knowledge. Your work will be showcased in front of this community

Recognition as a subject matter expert—Many authors have got an unparalleled recognition through their posts and work on Analytics Vidhya

Network with our community of over 2,80,000 data science and machine learning professionals and enthusiasts

Analyticsvidhya.com was registered on this day 4 years ago. In these 4 years, what started as a part time blog has transformed into one of the top data science portal across the globe. What started with an aim to make sure that Kunal (founder) learns regularly is now helping millions of people learn data science daily.

serve millions of data enthusiasts to learn analytics and have a global community of more than 50,000 registered users.

The first office setup was a single room with a few laptops and chairs. It was in the same apartment in which Kunal lived with his family. The company operated from there for 2 years.

To produce more leaders in data science, in the same year , launched very first AV apprentice program.

In July 2015, AV conducted its first hackathon – Predict the Megastar. This was a remarkable moment in the history of AV. The initial plan was to host the hackathon on the Discuss platform and ask participants to mail their solutions.

Analytics Vidhya was making a mark in the industry. The company started getting featured at various analytics conclaves in colleges, companies, TechStory and several other industry newsletters.

In 2016, Analytics Vidhya published a salary report in association with Jigsaw Academy. The report was featured by Economic Times, The Hindu, and few others.

Aim to provide as many resources as possible for learning analytics. These resources include:

Training and tutorials: Stuff to get you going and make you better in analytics and data science

Tips and tricks related to Data Science, Machine Learning, Business Analytics and Business tools

Hackathons – Real life industry problems being released in form of contests

Case studies: Case studies of problems and their analytical solutions

Interviews of Business Analytics & Business Intelligence leaders

This four are main functions of Analytics vidya. This organisation is started from a single person and explore to many partners.the idealogy behind this is to develop the data science and explore to entire people as it changes the environment of technology.

LOVELY PROFESSIONAL UNIVERSITY

Brief description of the work done

Predicting a trip duration isn't something that has not been though upon. With the use of Google maps API one can find the estimated time it would take to move between two points in the city. However, a detailed analysis of the factors affecting a trip between two points in a city can be very useful for accurate and robust prediction. Trip duration is not as simple as it seems. It is data dependent and is governed by a lot many factors apart from distance and speed. This research primarily focuses on the possible important factors that are used as attributes for the trip duration prediction in the New York City. This data can be used by taxi vendors for better services to the users. The research work not only uses a prediction model but also gives an in-depth analysis of the factors associated with the New York City taxi trips. A city like New York is expected to have various factors and variations with respect to the trip durations. The dataset used for training and testing purposes in multi-dimensional and requires a lot of pre-processing. This research work involves application of relevant machine learning algorithms such as linear regression, random forests, lasso regression and decision tree algorithms, random forest, logistics regression for completion of the task. The final algorithm used in this research
work is linear regression

Data analysis has traditionally been characterized by the trial and error approach – one that becomes impossible to use when there are significant and heterogeneous data sets in question. It is for this very reason that big data was criticized for being overhyped. The availability of more data is directly proportional to the difficulty of bringing in new predictive models that work accurately. Traditional statistical solutions are more focused on static analysis that is limited to the analysis of samples that are frozen in time. Enough, this could result in unreliable and inaccurate conclusion

methodology

The trip duration and analysis can be broadly divided into three working modules. These modules have been explained

Data Pre-processing

In order to prepare the data for the experiment, the data is cleaned to remove outliers. It involves removing the null

values, replacing missing values with dummy values, and changing the format of the date and time values so as to
utilize them in the algorithm. It also includes removal of outliers.

Feature extraction

In order extract the features, the distance between the Pickup and drop-off locations was found out using haversine and Manhattan distance formulae. The data points were clustered using mini batch k-means clustering. K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. It aims to partition observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the clusterataset is based on the 2016 NYC Yellow Cab trip record data made available in Big Query on Google Cloud Platform

The attribute features of the dataset
include:

- id - a unique identifier for each trip
- vendor_id - a code indicating the provider associated with the trip record
- pickup_datetime - date and time when the meter was engaged.
- dropoff_datetime - date and time when the meter was disengaged.
- passenger_count – driver entered value of number of people travelling in the taxi.
- pickup_latitude - the latitude where the meter was engaged.
- dropoff_longitude - the longitude where the meter was disengaged.
- dropoff_latitude - the latitude where the meter was disengaged.
- store_and_fwd_flag - This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip
- trip_duration - duration of the trip in seconds

4. Proposed Work

The machine learning model employed uses linear regression with the assistance of K means algorithm for trip duration prediction. The learning model uses K means clustering for features extraction on

the basis of the pick-up and drop off points. With the help of clustering, the model gets new features in the form of clusters. As many as 100 clusters were fetched and hence each cluster fetched two features which included a pickup cluster and a drop off cluster. These clusters accounted for 200 additional features which greatly helped in feature extraction and hence better prediction of the trip duration. Further, the linear regression algorithm was applied with the

varying parameters for actual final prediction. It not only provides an additional advantage of accuracy but also helps greatly in dealing with large datasets as in the case of NYC .taxi trip duration.

5. Experimental Research

The motive of this model is to find the best and most accurate way of predicting the trip duration using the existing features and adding some more features. The three most important steps in the experimental research which define the modules of the model in detail are:

1. Data Pre-processing

In this part, the major goal was to clean data in a way that the data is understandable and easy to predict. In order to achieve it, the following steps were followed-

 Elimination of outliers from the trip duration attribute in which the values lying outside the range of m-2*s and m+2*s were removed, where m is the mean of trip duration and s is the standard deviation of the trip duration.

 Eliminating the pick-up and drop off points that lie outside the New York City border.

```
print('The no of rows with distance =0 are {}'.format(len(df[df.distance==0])))
```

2. Feature extraction and enrichment:

This is the most important step before predicting the trip duration. In this step, the k-means clustering algorithm [7] was applied. It can be seen that there were many pick-up and drop-off points that were overlapping with each other, so to make computation faster k-means clustering was applied by making those redundant as one cluster A total of 100 clusters were formed using the pick-

up and drop-off points. The result of these yielded 100 clusters each producing one feature each for pick-up and drop-off location forming 200 .

```
df['pickup_day']=df['pickup_datetime'].dt.day_name()
df['dropoff_day']=df['dropoff_datetime'].dt.day_name()
df['pickup_month']=df['pickup_datetime'].dt.month
df['dropoff_month']=df['dropoff_datetime'].dt.month
```

Prediction

It was used for final prediction of the trip duration in the test dataset. The features were as many as 289 and the dataset was also very large, as a result for this type of problem linear regression was applied in which all the attributes were taken .

```
mean_pred=np.repeat(X_train[target_col].mean(),len(X_test[target_col]))
from sklearn.metrics import mean_squared_error as mae
sqrt(mae(X_test[target_col],mean_pred))
```

RESULT

The machine learning model that was used for the prediction task fetched a RMSLE of 0.44076. There
was little to no variation when the number of iterations for training was changed above 150. Another approach
which used linear regression without the usage of „New York City Taxi . The machine learning model not only fetched good accuracy but also gave an in depth analysis of the taxi rides variation over time. shows the variation
of average speed of taxi throughout the hours of the day, the day of the week, the month of the year.

Conclusion

The research problem was finding the trip duration a taxi takes for given pick-up and drop-off locations.
Trip duration was predicted using the linear regression algorithm. Mini-batch k-means was also applied for feature engineering and better performance in terms of both accuracy and computation times. Taxi giants such
as UBER and OLA can use the same data for analyzing the trends that vary throughout the day in the city. This not only helps in better transport analysis but also helps the concerned authorities in planning
traffic control and monitoring. The machine learning model used also helps in comprehending the important
factors that contribute in prediction of taxi trip duration .

The project can be done by using data exploration also which includes only data analytic but in linear regression we have included many formulas distance, square root and import of libraries like Sklearn , linear regression.
I had done this project because this helps the taxi company to plan their fleet in much better manner. So if we can predict what is the time when trip would end and where the position trip would end. So we can plan the fleet accordingly.

REFERENCE

1. ANAYLATICS VIDYA COURSE
2. https://medium.com/analytics-vidhya/building-a-linear-regression-model- on-the-new-york-taxi-trip-duration-dataset-using-python-2857027c54f3

THANKYOU