PREDICTING THE BEHAVIOUR OF A DRIVER BY USING TELEMATICS AND MACHINE LEARNING TECHNIQUES

Done by,
Madhavi Boyapati
2942211

Submitted in partial fulfillment for the degree of

Master of Science in Big Data Management and Analytics

Griffith College, Dublin, Ireland

June, 2020

Under the supervision of *Dr.Viacheslav Filonenko*

DISCLAIMER

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the Degree of Master of Science in Applied Digital Media at Griffith College Dublin, is entirely my own work and has not been submitted for assessment for an academic purpose at this or any other academic institution other than in partial fulfilment of the requirements of that stated above.

Name: **Madhavi Boyapati** Date: **12-June-2020**

ACKNOWLEDGMENTS

I owe my sincere gratitude and thanks to my professor and guide **Dr.Viacheslav Filonenko** for guiding me throughout the process of implementing and documenting. I also would like to acknowledge my friends and seniors support who guided me with the knowledge they possess. I would like to extend my warm gratitude towards my family and well-wishers who have been a great support morally and made me achieve and reach this stage.

ABSTRACT OF THE RESEARCH

For insurance companies it is very important to figure out how to make it more difficult to report a fraudulent claim and reward the drivers for good driving with low premium charges. Through deep learning and telematics techniques, fraud insurance detection can be predicted accurately. Insurers benefit as they can react quickly to their customers by using the data given by telematics-prepared vehicles. Insurers are in a better situation to assess the risk based on the data provided by telematics-equipped vehicles. The information permits insurers to offer 'pay-how you-drive'. Telematics recorded data will give the driver's behaviour with second by second basis, this will help for the insurers to find out more discrete driver's behaviours. From this, insurers can assess the data with a great accuracy whether the data given to them are correct or not.

Many drivers are opting to install black box technology, known as telematics. Telematics in vehicles uses real time data of the vehicle to monitor driving behaviour and figure out the premium accordingly. This requires a case with programming and sensors to be introduced in an individual's vehicle, and programming to process the information. Sensor technology will collect the speed, acceleration, odometer, fuel consumption etc. And the results will ultimately support the insurers and drivers through analysing the driver's behaviour.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

1.1 ML IN VEHICLE INSURANCE DOMAIN:

Insurance is intended to ensure the individuals and things we value most. Against risks and damages, auto insurance industry is providing financial security for the drivers. To gain benefit and attract the customers, driving behaviour of an individual and insurance pricing models are playing an important role **(He et al., 2018).**

Insurance companies are facing difficulties to find out fraudulent claims. Many drivers are opting to the technology called as telematics to their vehicles. Telematics in vehicles collects the real time data by monitoring the driver behaviour with time. This will result the more accurate data. The data allows insurers to offer 'pay-how-you-drive' models. To process the data, it requires a sensors and software to be installed in an individual's vehicle. The sensor technology collects the data like location, speed, acceleration, placement etc **(FRISS, 2020)**. Driving feedback of a customer will be provided by using recorded telemetry data of a vehicle (**Ayuso, Guillén and Pérez-Marín, 2020).**

1.2 PRIMARY GOALS:

The goal of the project is to use telematics data of the vehicles and analyse the behaviour of a driver whether it is good driving or not. As the current policies of insurance companies stimulate users to drive more kilometres yearly, does not reward the good driving behaviour and on the other hand it does not penalize the reckless driving. By analysing the telemetry data, which have collected from the vehicle, insurance companies are able to assess the risk at their best. When Customer claims by using the accident data given by telematics installed vehicle, the insurers can react quickly to their claims. Which results in customer satisfaction. It will reduce the cost of investigation and saves time **(Tselentis, Yannis and Vlahogianni, 2016).**

The customers will also get benefit from this telematics technology, as they will rewarded with low premium for safe driving. Telematics will give the immediate feedback about the risk and cost of the insurance. It will prevent the driver from careless driving. It enables the drivers with self-safety. This method is a motivation for the customers to improve their driving performance. And results decreasing the number of accidents in which somebody causes or engages in. By using this method will helps to improve the driving behaviour and thus emission of pollutants level decreases. This research project aims in building the deep learning model for implementing analysis of the driver's behaviour by using telemetry data of the vehicle. And the results will ultimately support the insurance companies through analysing the driver's behaviour **(Tselentis, Yannis and Vlahogianni, 2016).**

1.3 OVERVIEW OF APPROACH

- Extract the data set.
- Extract the data from json format to .csv.
- Pre-process the converted data.
- We can see the data for different lanes and turns.
- We can separate the data into training and testing.
- We will test the model through input data.
- We can see the driver behaviour.

1.4 RESEARCH QUESTIONS

- How far the telematics data can be used for analysing the driver details?
- How the machine learning solution can be used for predicting whether the driving was safe or harsh?
- What are the benefits for the vehicle insurance providers in analysing the telematics data before quoting the insurance policy value?

1.5 PROJECT STACKHOLDER

The stakeholders of this research implementation would be the insurance company and they could use the proposed machine learning model for analysing the telematics data generated from the IoT device or similar device of the vehicle and make a

prediction on the policy premium value. The result would be in the percentage, which will show how far the driving was harsh in terms of driving and helps to take up a decision.

CHAPTER 2. BACKGROUND

2.1 LITERATURE REVIEW:

The literature review part of the research work helps us to understand the related work that has been already done on the research work that we have chosen for implementation. This phase is considered to be one of the major part, as we can get to know the different approaches that can to be followed and on top of that, it will helps us to anticipate any issue that we may face in the implementation well in advance by analysing the different research works.

2.2 REVIEW WITH RELATED TO TELEMATICS:

Information has generally been one of protection industry's most important resources. device-to-device transmission is introducing the accident coverage industry with a Significant challenge **(Naic.org).** Back up plans are contributing on their capacity to gather, store, oversee and examine tremendous measures of variable information to tackle complex issues to stay productive. Accident protection is quick turning into a major information industry, with telematics based UBI ready to possibly change the matter of protection as we probably are aware it.

Based on the regularity and duration of journeys taken, informational indexes can illustrate up to 15MB of information every year, per customer. A safety net provider with 100,000 guaranteed vehicles can gather more than one terabyte of information for every year. The expense of the innovation and the equipment—as well as the roundabout expense for establishment, support and management—is one of the fundamental restricting elements to the faster and more extensive selection of telematics **(Handel et al., 2014)**. As the innovation gets less expensive, the versatility and accessibility of telematics-based protection programs is relied upon to develop at a quicker rate. The enormous information requests as far as capacity and examination, alongside the absence of normalization in telematics gadgets, present huge difficulties to safety net providers in their push to effectively incorporate telematics in their data innovation (IT) foundation.

The telematics gadgets for the most part utilized by insurance agencies are connected to the ready diagnostics port of a vehicle or some makers are integrating the system with vehicles. The sort of information recorded and transmitted from the vehicle shifts as indicated by the telematics innovation selected and by customers eagerness to share individual information. Sensors in telematics gadgets can catch information as basic as date, time, area furthermore, number of kilometres driven, speed, turnings, changing of lanes and acceleration. Right now, there are four classifications of telematics arrangements accessible in the market:

Dongle: the device is introduced by insurance company. You can install this by yourself and to be utilized for a certain time, generally for a half year. This device has the highest demand in America because of minimal price, re usability and more consistent quality. The driver can able to install this by self, can be moved to vehicle naturally turns on with the vehicle's start, creates high-calibre and secure information on the spot and driving behaviour, and can be packaged with other worth included administrations. **(Arumugam and Bhargavi, 2020)** alongside its numerous qualities, the dongle has various drawbacks, such as the way that it must be utilized in present day vehicles, is helpless against extortion as it could be altered since it can't be hard-wired into the vehicle's hardware, and will have short time be innovatively out of date.

Black Box: The expertly introduced black box, well known in Europe, is thought of to be one of the most assured and authentic. The black box can be utilized with PAYD and PHYD, and it can give the definite information about behaviour of ta driver. Since PHYD plans will in general be the generally modern of the telematics, UBI items require gadgets like the black box with coordinated accelerometers to follow an assortment of execution information like speed, gforces in hard turns and slowing down **(Handel et al., 2014).** The black box, notwithstanding its own sensors, can utilize the vehicle's inward sensors by connecting with its electronic control unit. The black box is additionally obviously appropriate for first notification of misfortune benefits as it is fixed in the vehicle undercarriage, giving early notification in case of burglary and significant data for scientific accident remaking on account of a mishap. The black box is additionally liked for following driving conduct information of youthful and unpractised drivers. And the price of this device is very high **(Park, 2020).**

Embedded: up to eleven vehicle manufacturers are inserted embedded telematics set up. While at an opportune time, inserted telematics gave administrations, for example, remote diagnostics, route and infotainment administrations, presently they can convey UBI administrations. The installed device associated with the vehicle's ECU can Capture and transmit an abundance of information about the vehicle's exhibition. It can improve the client relationship due to the advantages of this device. Some significance changes with implanted telematics are the similarly significant expense or the purchaser, absence of normalization, similarity with protection arrangements and out of date quality **(Handel et al., 2014)**.

Smart Mobiles: Phone media transmission innovation is the most recent device in telematics, with cell phones filling in as independent gadgets or connected to vehicles' frameworks to transfer a different type of data to and from the vehicle. Cell phones are a perfect telematics arrangement as they are regularly furnished with a large group of important sensors, such as GPS, accelerometers, and gyrators. They additionally have huge information stockpiling limit, or vast with the cloud, and prevalent correspondence abilities. The cost of this telematics tool is very less when compared to the remaining because it does not require any installations. However, the data collected by this device is not reliable as it requires a continuous adjustment **(Wahlstrom, Skog and Handel, 2017).**

2.3 REVIEW WITH RELATED TO INSURANCE INDUSTRY:

Insurance policies are provided to customers while they make a new purchase, which helps them to feel safe and secure in case there occurs any kind of accidental situations. According to **Ma *et al.* (2018, p. 246)**, Based on the parameters of using the existing policies are categorized into three different domains. These are paid how you drive (PAYHD), pay as you drive (PAYD). The third domain mile based is simpler than the other two as it considers the number of miles a vehicle is being driven. As opined by **Taherizadeh *et al.* (2018, p. 525)**, the major usage of telematics in this industry is to determine the major cause of an accident that is being taken place. Previously, it was very much complex to find out what caused an accident, which is a key measure to determine what kind of insurance will be applicable for a incident **(Sahebi and Nassiri, 2017, p. 66).**

To fit to modern living styles, insurance companies have gotten more customized. Anyway, technology has not generally kept pace. Increasingly more business is being finished over the internet, and in any event, utilizing cell phones. **(A. McCall, D. Bryant, W.S. Riney and E. Gay, 2020)**. A significant worry for insurance companies is to find out which one is a fraud claim among the claims. Many examinations have inferred that enormous rates of cases seem to include false claims. The insurance research council examined claims from nine states and found that 21 to 36 percent of the cases reckoned fake claims. **(Tennyson and Salsas-Forn, 2020)**. The Insurance Information Institute (2004) assessed property and loss (P&C) false claim at $31 billion of every 2002.

If not appropriately tended to, insurance misrepresentation not just puts the benefit of the guarantor in danger, yet additionally adversely influences its worth chain, the insurance industry, and might be very hindering to set up social and financial structures. Besides, all genuine policyholders are sufferers. Cheating is broadly accepted to build the cost of protection. This cost segment is borne straightforwardly by totally guaranteed parties as expanded premium rates. At long last, cheating speaks to a danger to the very standard of unity that keeps the idea of alive **(Viaene et al., 2007).**

2.4 OTHER RELATED RESEARCH PAPERS:

(Ayuso, Guillén and Pérez-Marín, 2020) has investigated that drivers who bought an insurance policy attached to a vehicle utilization. A GPS was introduced in their vehicles and the drivers were educated that their insurance would be corresponding to the quality of miles driven. Different measurements followed included scenes of extreme speed, urban driving and evening time driving, and these were additionally mulled over in the general rating framework. Here, we study the role played by occurrence and its effect on the danger of being associated with a mishap as far as the time and kilometres travelled by drivers under thirty. The data recorded by the GPS framework incorporates the aggregate number of kilometres voyaged and various factors catching components of the members' driving examples. This data was gathered

for various timespans during every year, as distinguished by the relating starting/end dates. The individual who involved in the accident is based on average distance travelled by the vehicle and experience of the driver.

(A. McCall, D. Bryant, W.S. Riney and E. Gay, 2020) in their work used a metadata to process insurance claims. Precise documents and productive checking with proofs is important for the insurers to identify the fraud claims and coverage. In Traditional insurance claim process physical inspection is needed to verify the claim by sending an agent. To avoid this, the insurer can simply verify the type of claim by analysing the vehicle electronic data. Metadata contains the details of the data like date, time, place, size of the data etc. for example, when a customer take a photo of a vehicle it the digital photo will provide the camera type, date of the photo was taken along with time, orientation of the camera details and location (GPS) where the photo was taken.

In auto insurance industry the metadata plays an important role. Suppose a customer met with an accident four after the lapsing of an insurance policy, then that individual went to reinstate the insurance policy on the very next day. And after that if that person claims an insurance by showing some digital photos or videos, then the insurers will verify the metadata of the files. If they don't match, then the insurers can conclude that it is a fraud claim **](A . McCall, D . Bryant, W.S. Riney and E . Gay, 2020)**. Customer will be rewarded for providing the digital photo of a vehicle when taking the new insurance policy or renewing the old one.

As of late, the fast improvement of telematics in insurance industry has empowered to gather enormous sums of fine-grained versatility information, like vehicle speed, quickening, motor speed, etc, to more likely profile drivers' hazard for estimating. With these telematics information, conventional techniques are normally utilized to process the protection cost, e.g., Pay-How-You-Drive model Despite the fact that the mass of new telematics information can possibly show driving practices all the more precisely and improve the granularity of hazard forecast, it likewise presents new research difficulties **(He et al., 2018).**

The author plan to address three different tasks. One is Driver Behaviour Profiling for Risk Forecast. To more likely model driver practices and dangers, we intertwine varying information, specifically vehicle telematics information gathered from the mainstream On-Board Diagnostic (OBD) gadgets, furthermore, UBI information that incorporates data, for example, the driver's segment, protection and guarantee reimbursement information. Trajectory based highlights are first removed to show the transient hazard change design for every person. To ease the information sparsity what's more, irregularity issue, propelled by the gathering level understanding, power-law-based outfit learning is performed to arrange drivers' diverse transient personal conduct standards inside gatherings, empowering increasingly dependable forecasts of driver hazard likelihood. customized estimating model that consolidates not just the segment examination, yet in addition the portability elements of the driver's hazard likelihood and voyaged mileage. The produced evaluating model is in this way versatile to individual hazard practices and can bolster different premium periods. Last one is Organization Profit Maximization. We likewise propose a useful answer for augmenting organization benefit, under the venture requirements.

To take care of the previously mentioned issues, we propose a bound together PPP (Profile-Price-Profit) system including three significant models. In particular, the Driver Behaviour Profiling Model examines direction-based driver hazard from OBD and UBI information, models worldly hazard change designs in time arrangement, and proposes a gathering level answer for foresee future hazard. The Insurance Pricing Model produces a driver's protection cost dependent on versatility factors separated from the driver conduct profiling model, and segment data. The Company Profit Model handles the benefit amplification issue, which is end up being NP-Complete, and gives a heuristic-based unique programming arrangement.

To anticipate a driver's future driving danger. we anticipate a driver's hazard from directions to catch the portability design. At that point, considering the time-differing property of directions, we segment the information by week, and fit the conduct in the time stream through the Power Law design. to conquer the difficulties of information sparsity and information irregularity in the hazard expectation, we find the gathering level knowledge, and propose an iterative refinement calculation

arrangement by group learning. In this paper, author proposed PPP system to enable the insurance agencies to give the customized protection cost what's more, accomplish the maximal benefit **(He et al., 2018)**. Initially, PPP fine-grained profiles the driver practices in time stream. In the meantime, a troupe learning calculation is proposed to foresee the driving danger by considering the gathering level understanding. At that point, PPP produces customized protection cost with adaptable premium periods. Both the driving conduct and the segment data are thought of. At long last, the maximal benefit issue is demonstrated to be NP-Complete and an obliged dynamic programming arrangement is proposed. PPP is assessed completely on the certifiable enormous scope OBD and UBI information. Exploratory outcomes shown that, PPP accomplishes close to the maximal benefit for the organization under this present reality requirements, brings down the aggregate cost for the drivers and is exceptionally commended by area specialists **(He et al., 2018).**

According to **(Ngai et al., 2011)**, the idea presents a methodology for fraud insurance detection in insurance business by applying different information mining strategies. At first, the most important parameters are selected from the actual dataset by utilizing a developmental calculation-based element determination strategy. The adequacy of the proposed framework is illustrated by leading a few tests on a true accident protection dataset. Also, a near investigation with another methodology legitimizes the predominance of the proposed framework.

An automobile insurance is a legitimately standing agreement marked between an insurance agency and proprietor of a vehicle (guaranteed) to offer money related help during vehicular burglary or harm. Collision protection extortion delineates a circumstance, where the guaranteed attains the monetary benefit by submitting manufactured reports to the organization by demonstrating harm to the vehicle in arranged mishaps or money related cases for past misfortunes **(Ngai et al., 2011).** This misrepresentation can be done by people, like, drivers, chiropractors, carport mechanics, legal advisors, cops, protection laborers furthermore, others. The insurance protection extortion can be isolated into various types, for example, recording a bogus protection guarantee document (a simpler way), to an increasingly beguiling way like manufacturing a mishap or auto burglaries. The proposed framework recommends a novel half and half extortion recognition procedure that attempts to recognize the fraud

cases from the ordinary ones from an under sampled adjusted dataset by utilizing an element determination strategy and an administered classifier. At first, the most pertinent parameters are chosen from the first imbalanced protection dataset by utilizing a transformative calculation-based element choice procedure. A test set is then separated from the chose quality rundown and the staying set is utilized for an under-sampling approach.

The author summarizes that Recognition of deceitful cases in accident coverage claims is an extremely challenging work as the case information is profoundly tilted in nature. In this examination, a novel methodology has been recommended that proposes the utilization of highlight determination strategy followed by grouping pair. This paper further researches the adequacy of applying a weighted ELM for the segregation of ill-conceived claims from the real ones.

(Šubelj, Furlan and Bajec, 2011) proposed a specialist framework for location, and resulting examination, of gatherings of teaming up accident coverage fraudsters. The framework is depicted and analysed in extraordinary detail, a few specialized difficulties in identifying extortion are likewise thought of, for it to be material by and by. Restricted to numerous different methodologies, the framework utilizes systems for portrayal of information. Systems are the most common portrayal of such a social area, permitting plan and investigation of complex relations between elements. False elements are found by utilizing a novel evaluation calculation, Iterative Assessment Algorithm. Other than characteristic qualities of substances, the calculation investigates likewise the relations between elements. The model was assessed and thoroughly investigated on certifiable information. Results show that collision protection extortion can be effectively identified with the proposed framework and that proper information portrayal is imperative.

Fake claims can be happened in different scenarios. It comes taking all things together variety shapes and sizes, from conventional misrepresentation, for example (basic) charge cheating, to progressively advanced, where whole gatherings of people are teaming up to submit misrepresentation. Such gatherings can be found in the insurance protection. To get money from the insurance companies, cheaters create a

fake mishap with their vehicles. They will frame accident that has never happened, and the vehicles have just been set onto the street. All things considered, most of such extortion isn't arranged (crafty misrepresentation) – an individual just takes advantage of the lucky break emerging from the mishap and issues overstated protection claims or on the other hand asserts for past harms. Most accidents have many similar things in usual. They frame the accidents in non-rural areas to minimize the number of observers. to make the scene easier the police is constantly called to the place. Most of the insurance cases are investigated by human rather than computer. As discussed so far, the proposed master framework utilizes systems of crashes to appoint doubt score to every substance. These scores are utilized for the identification of gatherings of fraudsters and their comparing crashes **(Bodaghi and Teimourpour, 2020).**

Author summarizes that, master framework approach for observation of categories of collision protection fraudsters with systems. Exact assessment shows that such misrepresentation can be effectively distinguished utilizing the recommendation and, specifically, that appropriate portrayal of information is indispensable. For the framework to be appropriate practically speaking, no marked informational index is utilized. The framework rather permits the ascription of special master, and it can in this way be embraced to new sorts of misrepresentation when they are taken note. The methodology can help the space agent to distinguish also, examine misrepresentation a lot quicker and all the more effectively. Besides, the utilized structure is anything but difficult to actualize and is too appropriate for identification (of misrepresentation) in other social areas **(Šubelj, Furlan and Bajec, 2011).**

(Burri and Burri, 2020) Has analysed the insurance claims by using machine learning techniques. According to reason behind the Machine learning (ML) to use in the insurance industry is it can deal with any type of data. For example, organized, unorganized, structured, and semi structured data. Risk assessment, risk identification, customer claims can predict the Machine Learning accurately. There are many advantages of using machine learning technique in insurance industry. detecting the fraud insurance claims, risk analysis, insurance claims are major things that machine learning perform. machine learning is not a new technology; it is there from the last few years. ML technology has three main learnings. They are supervised learning, unsupervised learning, and reinforcement learning.

Over the last few years, With the known variables in various sequences to obtain the required output supervised learning model was used. Currently insurance companies are showing interest to use the unsupervised methods. In this learning, if there are any changes in the parameters it automatically finds them and tries to modify with respect to the goals **(Burri and Burri, 2020).** The author discussed methods to alter the machine learning techniques into insurance business. They are automated and personalized item contributions, improved risk evaluation and upgraded false claim detection. To breakdown huge part of information and obtain the view of the customer actions. Customers permit insurance providers to go into basic product offerings with normal cost and quality. Insurance industry can offer customized items and arrangements which depend on the explicit necessities of small range. Companies get profit from receiving the significant development of customer activities. Machine learning predictions are very precise (**Roy and George, 2017).**

The summary of machine learning techniques in insurance industry yields that its analysis of insurance claims is done successfully. Currently, to make their views into different fields of business technologies are moving very quickly. In this regard, the insurance business does not need behind the others. In this manner, the way that insurance agencies are effectively utilizing information science examination is not astonishing. In pith, the point of applying information science examination in the protection is equivalent to in the other industries to advance showcasing methodologies, to improve the business, to upgrade the pay, and to lessen costs **(Burri and Burri, 2020)**.

(Baecke and Bocca, 2017) examination explores how this sensor information can improve the hazard choice procedure in an insurance agency. More explicitly, a few hazard evaluations models dependent on three distinct information mining methods are increased with driving conduct information gathered from In-Vehicle Data Recorders. This study demonstrates that including standard telematics factors fundamentally improves the hazard evaluation of clients. Subsequently, safety net providers will be better ready to tailor their items to the client's profile. Over the previous years, the exceptional count procedure for engine insurance agencies was mostly founded on general components. Vehicle explicit qualities and drivers' socio-demographical

information were the main contribution for the estimations. This from the earlier methodology can be improved by taking the case history of the customer into account. Using a legitimacy bad mark past cases-based model the back up plans could characterize a progressively reliable degree of peril. Despite the wide selection of such models, they despite everything have constraints in assessing the genuine hazard level of the policyholder. More explicitly, the presentation to the hazard is not yet thought of.

These days, under the expanding rivalry, attempting to accomplish a cost decrease for both the safety net provider and the customer, some insurance agencies have created Usage-Based-Insurance (UBI) models. Through the utilization of In-Vehicle Data Records, the safety net provider can gather driving conduct information of every client. These records incorporate the kilometres driven, partitioned dependent on the spot and time. Insurance agencies can increase a solid upper hand by effectively utilizing and breaking down this information. By using this information insurers are able assess the risk effectively. They were used machine learning algorithms such as logistic regression, random forecasts, artificial neural networks. Then build a k fold cross validation to find predictive models of the ML models.

The author **(Baecke and Bocca, 2017)** concludes that examination has explored the effect of these information on the risk determination process. More explicitly, it is the main investigation that demonstrates in detail the included prescient estimation of telematics information not withstanding generally utilized factors, for example, client explicit, vehicle explicit and recorded cases factors. A prescient model that is just founded on this information source is now ready to evaluate the mishap hazard superior to conventional models. Be that as it may, most worth untruths in consolidating the two information sources since they catch distinctive basic components of the hazard. Insurance agencies ought to animate their customers to introduce In-Vehicle Data Recorders. This can create points of interest for the two safety net providers and clients. While this plainly improves a backup plan's hazard choice procedure, clients can profit by a lower premium if their driving conduct is breaking down as protected, yet additionally from extra administrations, for example, programmed crisis calls, taken vehicle following and symptomatic administrations.

(Ayuso, Guillen and Nielsen, 2020) showed how information gathered from a GPS gadget can be consolidated in engine protection ratemaking. The count of premium costs dependent on driver behaviour represents a good choice for the protection segment. The methodology depends on check information relapse models for recurrence, where introduction is driven by the separation voyaged and extra parameters that catch qualities of vehicle use and which may affect asserting conduct. He proposed actualizing an old-style recurrence model that is refreshed with telemetric data. We outline the strategy utilizing genuine information from use-based protection approaches. Results show that the separation went by the driver, yet in addition driver propensities, significantly influence the normal number of mishaps and, subsequently, the expense of protection inclusion. This paper gives an approach including a progress estimating moving information and experience that the organization previously had before the telematics information showed up to the new world including telematics data.

Telematics is the innovation of sending, accepting, and putting away data by means of media transmission gadgets related to affecting control on remote articles. In this manner, vehicle telematics permits driver data to be gathered utilizing an electronic gadget. Comprehensively talking, this GPS-based innovation records mileage notwithstanding other information identified with driver conduct. Old style protection ratemaking depends on recurrence and seriousness models that foresee the normal number of cases and their normal expense on the grounds of verifiable data put away in an insurance agency's database. Generally, the factors remembered for the prescient models are gathered about the driver and vehicle at the hour of strategy issuance, however data about driving propensities are not considered straightforwardly because driving style and force couldn't up to this point be estimated impartially. This paper is especially worried about the change procedure from old style protection evaluating to protection estimating including telematics. Let us state an insurance agency needs to present telematics. Furthermore, let us state that this organization has a long history of understanding their clients and estimating their hazard. It presumably would not be a smart thought to toss away the authentic information and scholarly advancement the organization has acquired over a long time. A superior methodology is by all accounts to consider the issue as a three-phase process: first one is evaluating before telematics is presented, second one is the change to valuing including telematics, what's more, another system, where telematics information is completely coordinated in the business

forms of the organization. Thusly, in this paper we envision telematics to be acquainted with the insurance agency as an amendment to their present valuing.

The summary of this paper **(Ayuso, Guillen and Nielsen, 2020)** is in any case, the data gave by telemetry speaks to a significant change in the customary valuing framework, since dynamic data about the driver opens. This data incorporates not just the separations driven during a given timeframe, yet in addition the drivers' propensities and conduct that may experience changes during this time and which, thus, may be influenced using different premium rates. The consideration of mileage in the model methods genuine hazard introduction can be considered and, thus, actuarial premiums at the individual level can be more precisely determined**.**

According to **(A. Brandmaier, Gillespie and Hughes, 2020)** A correspondence module may get interchanges from a vehicle. A crash location module may discover that a crash has happened at a vehicle dependent on at least one of the correspondences got from a vehicle. A member identification module may recognize at least one member engaged with the impact dependent on at least one communication got from the vehicle. A deficiency assurance module may recognize one of the members as the to blame member considering an examination of a deficiency assurance ruleset to vehicle telematics information remembered for at least one of the communications got from the vehicle. A case handling module may decide if to document a protection guarantee related with one of the members dependent on the assessed fix cost.

The summary of this **(A. Brandmaier, Gillespie and Hughes, 2020)** yields that A correspondence module may get information from a vehicle or sensors connected to a vehicle. An accident recognition module may establish that a crash has happened at a vehicle dependent on the at least one of the information gotten from a vehicle or sensors connected to a vehicle. A member ID module may distinguish at least one member engaged with the impact dependent on at least one Data got from the vehicle or sensors connected to the vehicle. A flaw assurance module may distinguish one of the members to blame member considering an examination of a deficiency assurance ruleset to vehicle telematics information added for at least one of the communications got from the vehicle or sensors appended to the vehicle. A fix price evaluation component may figure an evaluated fix cost dependent on vehicle symptomatic information

remembered for at least one of the correspondences got from the vehicle or sensors connected to the vehicle. A cases expert cessing module may decide if to record a protection guarantee related with one of the members dependent on the evaluated fix cost.

The author **(Pesantez-Narvaez, Guillen and Alcañiz, 2019)** has proposed two models to predict the number of happening accidents. They are logistic regression and XGBoost. And compared the two models. Logistic regression can able to calculate the most extreme probability. Along these lines, the thought fundamental a calculated relapse model is that there must be a direct mix of hazard factors that is identified with the likelihood of watching an occasion. according to the author it is a most used model to predict the wanted output. Xgboost model has a high prediction occurrence when compared to logistic regression in training model. But, in testing data the performance is very poor. Summary yields that, the occurrence of accident claims was predicted with the logistic regression model and performance of this model is also very high.

According to **(Arumugam and Bhargavi, 2020)** A mishap is explained as an unlucky event that happens out of the blue and accidentally, ordinarily bringing about harm or injury. Taking all the results that could eventuate after a mishap into consideration, there are motivations to accept that an ordinary individual does not drive with an ex-bet aim to cause a mishap. Holding a legitimate driving permit is an essential to drive in any place of the world and during the authorizing procedure, individuals are taught about the driving standards and wellbeing measures to be followed. Despite every one of these, mishaps occur and shockingly, human factor is ascribed to be the principal reason causing the mishaps. Reasons, for example, interruption, tipsiness, speeding, running red lights and stop signs, carelessness, street wrath, forcefulness and laziness are positioned among the highest human elements.

The common explanations behind mishaps are ordered into three classifications: Bad climate or terrible foundation, vehicle failing (producing deformities or mileage) or human elements (physiological or social). While the physiological errors are going on because of driver exhaustion, tiredness, conduct slip-ups could take numerous structures, for example, occupied driving, alcoholic driving, forceful driving, street rage, hard increasing speed, hard slowing down and turning and speeding. Forceful

driving and street rage are from the earlier practices that are conceivably prompting lethal or non-deadly street mishaps, occurrences of physical savagery and even killings.

Rash driving includes driving the engine vehicle in a dangerous and unfriendly way without caring for other people, which remembers perilous conduct for street, for example, making successive or hazardous path changes, running red lights and stop signs, incorrect way driving, ill-advised turns, closely following, disregarding traffic controls. Street rage is an irate driving conduct showed by the driver, which incorporates making inconsiderate motions, making physical and verbal dangers, and displaying hazardous driving techniques focused towards another driver with an end goal to threaten or discharge disappointment. with number of accidents increases the number of insurance claims by the customers also increases. To get rid of this problem insurance companies are introduced usage-based insurance which leads to proper risk assessment.

(Arumugam and Bhargavi, 2020) proposed a model to detect the anomaly by using machine learning and big data techniques. the output of this method will provide the awareness to drivers for better protection during mishaps of rash driving.

CHAPTER 3. METHODOLOGY

3.1 CRISP-DM:

The data mining project follows several methodologies in developing the project from obtaining the dataset to visualizing the result. This research project adopts the CRISP DM methodology for building the data science project. CRISP-DM stands for Cross Industry Standard Process for Data Mining **(Hipp, 2020).**

The different stages of CRISP-DM methodology are listed below**:**

- Business understanding
- Data understanding
- Data preparation
- Modelling
- Evaluation
- Deployment

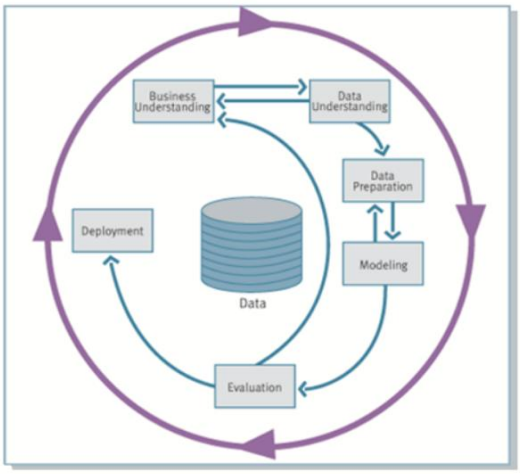

Figure 1: CRISP-DM

3.2 BUSINESS UNDERSTANDING:

This phase essentially implies what we are attempting to resolve and what is the client need or issue that we are attempting to explain, and afterward relying upon the sort of the issue that we are attempting to comprehend, at that point we can perceive what sort of arrangement we can use to take care of the issue. This phase involves finding the type of problem, in this research project the problem is identifying the behaviour of the driver. After that we can make a proper step to figure out the issue.

when we have distinguished the issue that we need to resolve by means of Machine learning. It must be referenced here that not all issues should be settled by means of ML, henceforth we have to observe cautiously what the business needs is here in this circumstance and we have to represent the entirety of the suspicions, objectives just as requirements also, that this business issue accompanies so we can locate an answer for this.

3.3 DATA UNDERSTANDING:

In this process the acquisition of data is more important to build efficient machine learning algorithms. In this phase we must perceive what sort of information that we have. We must see from where the information is coming from. In this project data is collected from DOI: http://dx.doi.org/10.5281/zenodo.1009540. The data is in large blocks of json files. The data having a large amount of details like speed, acceleration, fuel capacity, fuel consumption, etc. the data contains a five different directions namely left lane change, right lane change, left turn, right turn, straight **(Tonutti, Ruffaldi, Cattaneo and Avizzano, 2019).**

There are numerous measurable properties that we can watch that may add the synopsis measurements too, as though there are values that are clear or nulls and how would we have to deal with these, and furthermore if any anomalies and how would we have to deal with these. We can likewise perceive how all the parameters are associated with one another to observe its importance to remaining parameters and to check whether there a solid connection between certain parameters.

3.4 DATA PREPARATION:

The data preparation is one of the stages in the methodology process and is also the important stage where the acquired data is transformed into pure quality for the sake of building better machine learning models. The extracted data needs to be pre-processed to understand by machine learning algorithms for classifying the data into specific category with better accuracy. The data is stored in the json format.

The data stored in json format is converted into CSV file with python programming language and then concatenated all the CSV file to make final CSV. Furthermore, the converted data is then processed with several steps such as checking the null values, missing values, changing the categorical variables, adding the attributes to make a final dataset. Train the half of the data into safe driving and remaining half into unsafe driving.

Hereby, all the above-mentioned preparation steps are accomplished on the data. the extracted data split into five categories namely left lane change, right lane change, left turn, right turn and straight. All the steps are performed on the extracted data with the purpose of building efficient machine learning models.

3.5 DATA MODELLING:

Selecting models which is going to be used in the research project is the first step in the data modelling process. Selecting the suitable model includes a review of literature review and finding the commonly used predictive techniques which are already been successful. Based on the related work, the supervised machine learning algorithm is used in this research project. Classification method will predict the qualitative targets and regression technique predict the quantitative variable. In this project the target variable is qualitative one. So, Driver behaviour is coming under classification algorithm. To receive the data from a user as input an interface is created by using python programming language. This interface will generate the output.

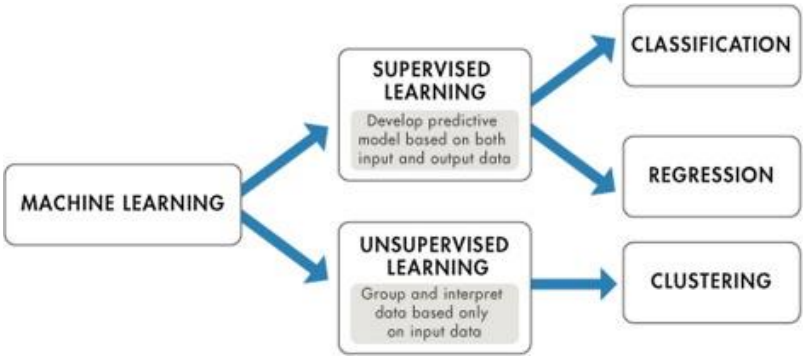

Figure 2: Supervised Vs Unsupervised Learning

3.6 DEPLOYMENT

In the last phase we deploy the model. It includes all the necessary code modules for the project to execute in a single package which is readily available and can be easily accessed by other developers. The developed code modules for this project is published in the cloud environment or server environment by stacking the code modules into a compressed packaged which is generally referred to as deploying the code in the release mode. The code modules for this project is developed using the python platform called PyCharm which is the Integrated development environment (IDE) and coded in python programming language. The code modules published in the cloud environment is easily accessible and hence it can be deployed within the local system to run the code with an ease.

CHAPTER 4. IMPLEMENTATION

In the implementation chapter, we will be discussing in detail on the machine learning models that we have used in our research work for the implementation. On top of that, we will be explaining about the user interface of the application that we have developed using the Python flask web framework. We have also mentioned the list of the packages that we have used in the Python programming and the purpose of using the same in our application.

4.1 MACHINE LEARNING MODEL:

The machine learning programming is used for developing a model that can understand the pattern among the data and make the decision with respect to the user input. In our case, the machine learning model will be trained with the dataset that has the telematics data generated from the heavy vehicles with the IoT devices and make a prediction on the how the driver was driving by classifying the findings into safe driving, average driving, and rash driving. To implement the discussed research, we have taken the consideration of three machine learning models which are listed below.

- Logistic Regression
- K-Nearest Neighbour (KNN)
- Long-Short Term Memory (LSTM)

The machine learning models has been chosen on based of the literature review papers that we have analysed in the literature review chapter. In terms of making the model to learn, we have used the supervised learning in which we train the model by labelling the telematics training data by having the classification of safe driving and unsafe driving. In the supervised learning, the model will be trained with the pre-processed training input and find the prediction on whether we would like to find how the inputted driving handled. We know the fact that no machine learning model is best and the type of data, the volume of the data and the other related factors decides which machine learning algorithm best fits for our case. So, it is essential to take multiple machine learning models for a single research work implementation and proceed with

doing the enhancements and deployment with the one that has shown the best accuracy and performance.

LOGISTIC REGRESSION:

Logistic regression is a supervised learning machine learning model calculation used to anticipate the likelihood of an objective variable. The idea of target or independent parameter is binary, which implies there would be just two potential classes **(Peng, Lee and Ingersoll, 2002).**

In basic words, the reliant variable is quantitative in nature having information coded as either 1 (represents achievement/yes) or 0 (represents disappointment/no). Numerically, a logistic regression model predicts P(Y=1) as an element of X. It is one of the least complex ML calculations that can be utilized for different classification issues. The following are the types of Logistic regression (Hoffman, 2019).

Binary or Binomial:

In this type the target variable must have two possible outcomes either 1/0 or yeas/no or true/false.

Multinomial:

In this type the dependent variable can at least have three or more. These can be unordered.

Ordinal:

In this type the dependent variable can at least have three or more. These can be ordered.

Assumptions of Logistic Regression:

- The predicted parameters must be always in binary format and the output should be denoted by 1.
- All the independent variables must be independent with one another.
- We should remember important factors for our model.
- We should choose a large sample size for logistic regression.

K-NEAREST NEIGHBOUR (KNN):

K-closest neighbours (KNN) model is a kind of supervised ML calculation which can be utilized for both classification and regression. KNN method processes the separation between each preparation test and test samples in the dataset and afterward returns k nearest tests. It is ensured to discover careful k closest neighbours **(Sun and Huang, 2010).**

The two properties would characterize KNN well –

Lazy learning algorithm− KNN is an apathetic learning calculation since it doesn't have a preparing stage and uses all the information for preparing while classification.

Non-parametric learning algorithm− KNN is likewise a non-parametric learning calculation since it doesn't expect anything about the fundamental information.

Working of KNN Algorithm

K-closest neighbours (KNN) calculation utilizes 'feature similarity' to anticipate the estimations of new datapoints which further implies that the new information point will be doled out a worth dependent on how intently it coordinates the focuses in the training set. Below steps explains the working of KNN **(Deng et al., 2016).**

a. For applying any model, we require a dataset. So, during the initial step of KNN, we should load the training information just as test information.
b. Next, we must pick the value of K for example the closest data points. K can be any number.
c. For each point in the test information do the accompanying –
    - Calculate the separation between test information and each line of training data with the assistance of any of the technique to be specific: Euclidean, Manhattan or Hamming separation. The most generally utilized strategy is Euclidean.
    - Sort them from smallest to largest based on the value of distance.
    - Now, it will pick the top K columns from the sorted cluster.

- Now, it will allot a class to the test point dependent on most regular class of these lines.

LONG-SHORT TERM MEMORY (LSTM)

It is uncommon sort of recurrent neural system that is equipped for learning long haul conditions in information. This is accomplished because the repetitive module of the model has a mix of four layers cooperating with one another. To classify, process and predict LSTM is most suitable one.

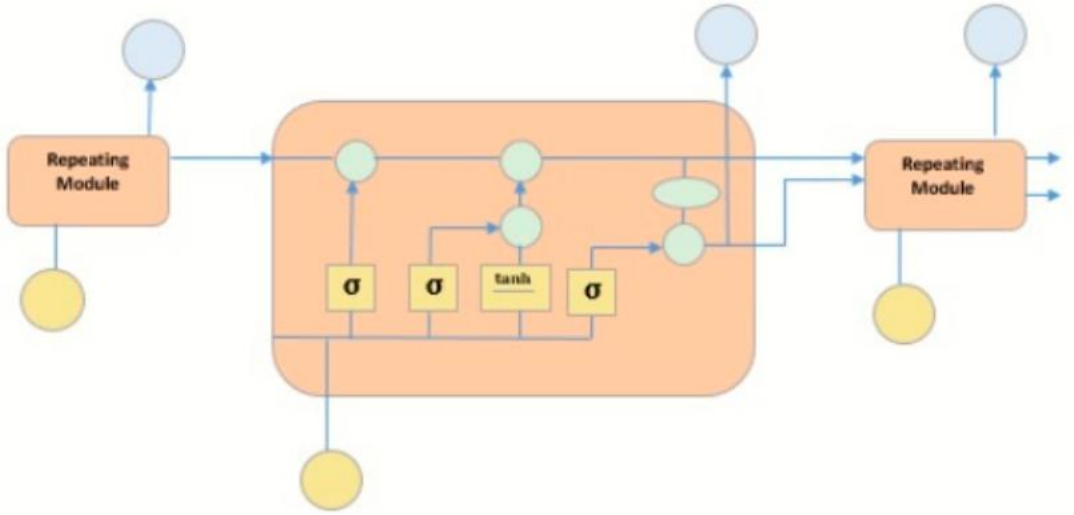

Figure 3: Long-Short Term Memory

As a part of implementation, we have started with pre-processing the input training data. In pre-processing, we have fetched the training records with respect to the labels, which is 0 and 1. Zero represents the safe driving and one represents the unsafe driving. Once the data split is done, the values are loaded into the data frame and calculated matrix is stored as a file which will be processed by the training code **(Gers, 1999).**

In the training code, we read the pre-processed data frames and load the same for further processing. As a next step, the data values are split into training dataset and test dataset in the ratio of 70 percent for training and 30 percent for testing. In the next step, we have created an LSTM model with all the layers and choosing the optimisers. Then we have fit the training and test data in the model and set the epochs to 800 to

increase the accuracy of the model. Once the training part is completed, we will be storing the trained model and use it later purpose.

4.2 WEB-BASED USER INTERFACE:

The research project has developed a web base user interface for the user to interact with the developed applications which makes the developed project as an interactive project. The web application is also developed using the python programming language and the python flask web framework is utilized in building the user interface. The user interface has option to enter the vehicle start time, as the dataset has only the seconds information that got captured from the installed device for the given trip. The calculations will be done on the telematics data with the seconds considering the user input time as a start time.

The actual data that we input to the model is the telematics data that got captured from the vehicle to have a check on how the vehicle was handled. The input data will be uploaded using the file input box option in the user interface and the same will be stored in a drive for processing. The pre-processing step will be applied by the code and the trained model will be unpicked and used for prediction.

The model will generate the following results at the end of the execution.

- Start Time (User entered)
- End Time (Calculated from the seconds field in the input data)
- Total Driving Time (Calculated field)
- Maximum Speed (Calculated field)
- Predicted Behavior (ML Outcome)
- Prediction Score (ML Outcome)
- LChange, RChange, Straigh, LTurn & RTurn with timestamp details (Calculated report)

Find the snapshot of the user interface below:

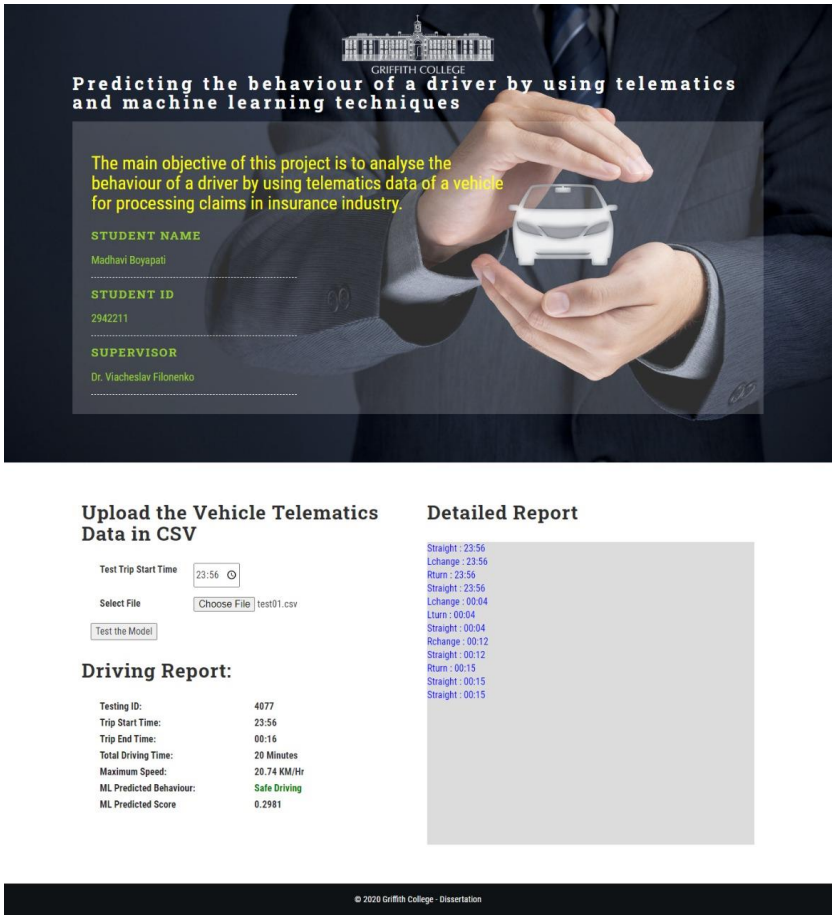

Figure 4: Web User Interface using Python Flask

CHAPTER 5. SYSTEM DESIGN SPECIFICATIONS

In this chapter, we will be elaborating the system configuration and specification details that we have used for developing the machine learning model.

Operating System: Windows 10 Operating System

Processor: Core i5

Programming Language: We have used python programming for developing the application. The python language is predominantly used for development of project in extracting the data, performing pre-processing on the extracted data, building the machine learning algorithms, performing analysis on the telematics data, developing the web based user interface and python based web framework called python flask is used for integrating the developed application **(Marwa, 2018).**

Python Packages: The names of the python programming package and the libraries used in the above-mentioned process are provided here.

Flask package is used for developing the web-based user interface for the project. The web application interface can be developed using the HTTP requests, JavaScript technology along with the python programming language.

OS Library: Very often it is required to interact with the operating system for training the machine learning model, perform the prediction and so on. The OS library package in python acts as an interface to the functionalities depends on the operating system and perform the required operation **(McKinney, n.d.).**

Pickle Library: In most of the cases, the trained model weights need to serialize and stored in a file or any other format in a system, so that the trained model can be used for performing the prediction without training the model instantly for every predictions. To do this, we need to serialize the model weight, which can be perform using the functions in the pickle library. It converts the data into character stream and the same can be deserialized into python object using the scripts **(Fasnacht, 2018).**

Math Library: The python Math library provides access to the common math functions that can be used to perform complex mathematical calculations and access to the constants in python. We do not have to explicitly install this library, as it is included by default in the python module **(Bergstra and Breuleux, 2020).**

Pandas: Pandas is a Python package that is used to perform analysis and manipulation on the data. It holds the data like in the excel file and it is called as Data Frames. The pandas rely on some of the other packages like Numpy and MatPlotLib **(McKinney, n.d.).**

Numpy: Numpy is a python package that is used to perform the scientific calculations. The multi-dimensional container of data can be maintained with the help of numpy package. Lists is a package that is like numpy and can be used as an alternative **(McKinney, n.d.).**

CHAPTER 6. TESTING AND EVALUATION

In this chapter, we will be discussing on the test executions with the machine learning model that we have build and the results we have obtained at the end of the execution. On top of that, we will be discussing on the evaluation phase that we have undergone for the implemented model.

6.1 TESTING:

Here we will discuss on the some of the test executions and the results that we have observed.

Test 01: We have given inputted the telematics data for a trip and our machine learning model has predicted it is a safe driving. On top of that, the model also has logic to calculate the maximum speed, and the total driving time. The average of speed observed and the seconds on the input data used for calculating the other factors given in the below report.

Detailed report column helped to understand the operations like turning into left, turning into right, driving straight, changing the lanes to left and right. This report will help the stakeholders to understand better on how the driver is performing the driving.

Threshold Values:

- If the model predicted score is less than .30, we can say then the driving was safe driving.
- If the predicted score is between .30 to .40, then we could say it is an average driving.
- If the predicted score is above .40, then we can say it is a rash driving.

Note: The above given threshold value is completely specific to the business and it may vary from clients to clients.

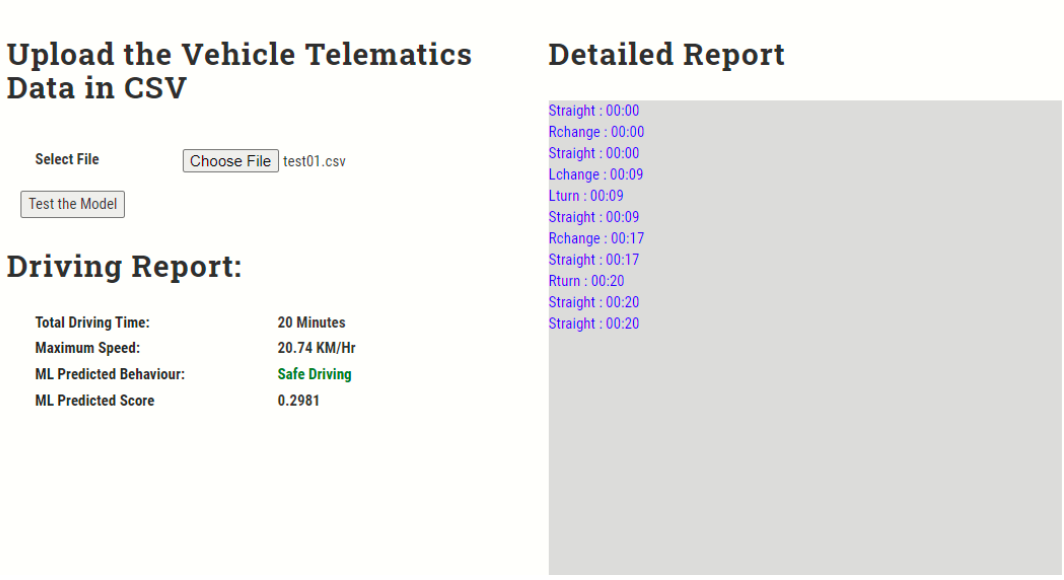

Figure 5: Test Execution 01

Test 02:

The below testing is executed to demonstrated to show the prediction of harsh driving.

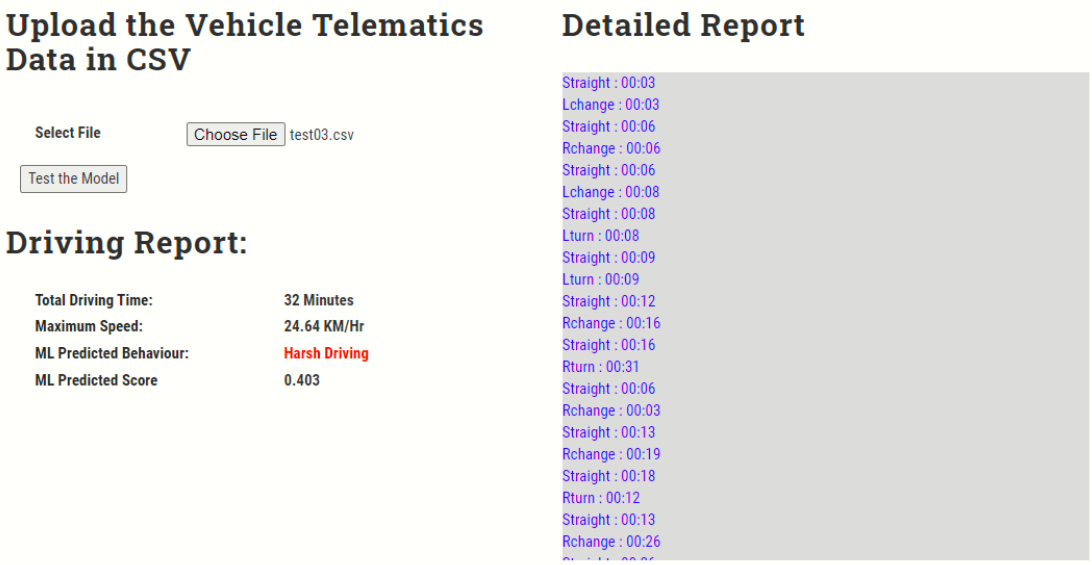

Figure 6: Test Execution 02

6.2 EVALUATION:

In this phase evaluation metrics needs to be observed to evaluate the model. In order to observe the performance of a model evaluation metrics are utilized. Model Performance measurements expect to segregate among the model outcomes. There are various metrics available to evaluate machine learning models. They are

Confusion Matrix:

It is a table that traces various forecasts and test outcomes and stands out them from true qualities. They are utilized in insights, information mining, ML models and other (AI) applications. Generally, confusion matrix is utilized for inside and out examination of information proficiently and quicker investigation by utilizing information representation.

- **Accuracy:** The total number of correct predictions.
- **Precision:** The number of positive instances that were effectively distinguished.
- **Negative Predictive Value:** The number of negative instances that were effectively distinguished.
- **Recall:** The amount of true positive instances that were effectively distinguished.
- **Specificity:** The number of negative instances that were effectively distinguished.

F1 Score:

The mean of precision and recall is considered as F1 score. The higher the score higher the prediction performance vice versa.

AUG-ROG (Area under ROG curve):

It is a chart which shows the performance of a classification algorithm. It plots two variables namely true positive rate and false positive rate.

True Positive Rate:

Number of True positive separated by the aggregate of the quantity of True positive and the quantity of bogus negatives. It depicts how great the model is at anticipating the positive class when the real result is certain.

False Positive Rate:

The number of wrong positives separated by the aggregate of the quantity of wrong positives and the quantity of genuine negatives.

ROG plot the curve between True positive rate and wrong positive rate. These plots are created at various grouping thresholds. So, if we have a low characterization edge, at that point we can ready to arrange more things as positive consequently expanding both False Positives and True Positives.

CHAPTER 7. CONCLUSION & FUTURE WORKS

In this chapter, we will conclude our research work by stating the complete path that we have travelled in making this research work implementation to get complete. On top of that, we know well any software application that needs to be enhanced to keep up the application standard matching with the latest needs. In this chapter, we will also discuss on the work that needs to be done on the completed research as a part of enhancements.

7.1 CONCLUSION

We have started our research from the business understanding phase. It was great learning when we have analysed about the problems that exists in the different domains and thinking about a machine learning solution that could address that problem and facilitate the business users to take an efficient decision. With the guidance of the professor, we have taken one of the specific problems that the vehicle insurance industry is facing in terms of setting up the policy value. No all vehicles are handled in the same manner, so it is required to understand on how the particular vehicle is handled in the part, to decide on the policy rate that we can set for the vehicle in the future. In this case, we have understood that the machine learning model would be an ideal solution that will analyse the driving records of the vehicle in the part and suggest the policy rate accordingly.

We have taken different algorithms for the implementation and this phase helped us to demonstrate the knowledge that we have gained through out the course of education and develop a solution that could solve a real-time problem. Evaluating and choosing the model, helped us to learn on how to compare any model that got developed and chosen the right one. We have chosen the LSTM model by comparing the same with the other two models by considering the different factors like the accuracy of the model and the performance as well.

On top of all, the whole process helped to learn the development phases and undergoing all the phases in the right way. The process also helped us to understand the

fact that any software development needs to undergo the software development process to create a solution that was expected.

7.2 FUTURE WORKS:

In the future work, we will be discussing on the updates that needs to be performed on the research work that we have developed for the academic purpose. This part of the document will help us to document on the direction that we need to take in the further phases of the development.

In the current version, we have developed only the machine learning model that denotes whether the driving was safe driving or not. And from the calculating value, the insurance company must manually fix up the insurance policy proposal. To over this manual work, in the further phase, a layer can be built on top of this machine learning model and purpose of the new layer would be suggesting the price that needs to be quoted for the given vehicle insurance. On the other hand, logically speaking, the threshold that will be considered on the rash driving score will be varying from one insurance provider to another insurance provider. So, to build this machine learning model as a software package, we must think about providing the configurable options on the threshold with respect to the insurance amount calculation and provide the application for the different insurance providers.

The futured work mentioned here will help us to end-up in building a software product as a service (SaaS) and make it available for use to the wide range of users.

REFERENCES

1. FRISS. 2020. How Telematics Reduces Car Insurance Fraud - FRISS. [online] Available at: <https://www.friss.com/press/how-telematics-reduces-car-insurance-fraud/>.
2. 2020.[online]Availableat: <https://www.researchgate.net/publication/304530360_Innovative_Insurance_Schemes_Pay_ashow_You_Drive> .
3. He, B., Zhang, D., Liu, S., Liu, H., Han, D. and Ni, L., 2018. Profiling Driver Behavior for Personalized Insurance Pricing and Maximal Profit. 2018 IEEE International Conference on Big Data (Big Data).
4. Time and distance to first accident and driving patterns of young drivers with pay-as-you-drive insurance.
   In-text: (Ayuso, Guillén and Pérez-Marín, 2020)
   Your Bibliography: Ayuso, M., Guillén, M. and Pérez-Marín, A., 2020. Time and Distance to First Accident and Driving Patterns Of Young Drivers With Pay-As-You-Drive Insurance.
5. **In-text:** (2020) A . McCall, T., D . Bryant, W., W.S. Riney, J. and E . Gay, C., 2020. METHOD FOR USING ELECTRONIC METADATA TO VERIFY INSURANCE CLAIMS. [online] Patentimages.storage.googleapis.com. Available at: <https://patentimages.storage.googleapis.com/b4/87/8b/fe92312af2bf50/US9818157.pdf>.
6. Tennyson, S. and Salsas-Forn, P., 2020. *Claims Auditing In Automobile Insurance: Fraud Detection And Deterrence Objectives*.
7. Iit.cnr.it. 2020. [online] Available at: <https://www.iit.cnr.it/sites/default/files/human-behavior-characterization.pdf>.

8. Viaene, S., Ayuso, M., Guillen, M., Van Gheel, D. and Dedene, G., 2007. Strategies for detecting fraudulent claims in the automobile insurance industry. *European Journal of Operational Research*, 176(1), pp.565-583.

9. Burri, R. and Burri, R., 2020. *Insurance Claim Analysis Using Machine Learning Algorithms*. [online] Ijitee.org. Available at: <https://www.ijitee.org/wp-content/uploads/papers/v8i6s4/F11180486S419.pdf>.

10. Roy, R. and George, K., 2017. Detecting insurance claims fraud using machine learning techniques. *2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT),*.

11. Ngai, E., Hu, Y., Wong, Y., Chen, Y. and Sun, X., 2011. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3), pp.559-569.

12. Šubelj, L., Furlan, Š. and Bajec, M., 2011. An expert system for detecting automobile insurance fraud using social network analysis. *Expert Systems with Applications*, 38(1), pp.1039-1052.

13. Bodaghi, A. and Teimourpour, B., 2020. *Automobile Insurance Fraud Detection Using Social Network Analysis*.

14. Patents.google.com. 2020. *US20060212195A1 - Vehicle Data Recorder And TelematicDevice-GooglePatents*.[online]Availableat: <https://patents.google.com/patent/US20060212195A1/en>.

15. Baecke, P. and Bocca, L., 2017. The value of vehicle telematics data in insurance risk selection processes. *Decision Support Systems*, 98, pp.69-79..

16. Ayuso, M., Guillen, M. and Nielsen, J., 2020. *Improving Automobile Insurance Ratemaking Using Telematics: Incorporating Mileage And Driver Behaviour Data*.

17. A. Brandmaier, J., Gillespie, J. and Hughes, S., 2020. *US8799034B1 - Automated Accident Detection, Fault Attribution, And Claims Processing - Google Patents*. [online]Patents.google.com.Availableat: <https://patents.google.com/patent/US8799034B1/en>.

18. Naic.org.2020.[online]Availableat: <https://www.naic.org/documents/cipr_study_150324_usage_based_insurance_and_vehicle_telematics_study_series.pdf>.

19. Handel, P., Skog, I., Wahlstrom, J., Bonawiede, F., Welch, R., Ohlsson, J. and Ohlsson, M., 2014. Insurance Telematics: Opportunities and Challenges with the Smartphone Solution. *IEEE Intelligent Transportation Systems Magazine*, 6(4), pp.57-70.

20. Park, J., 2020. *A Study Of Using The Car's Black Box To Generate Real-Time ForensicData*.[online]Koreascience.or.kr.Availableat: <http://www.koreascience.or.kr/article/JAKO200810737032965.page>.

21. Wahlstrom, J., Skog, I. and Handel, P., 2017. Smartphone-Based Vehicle Telematics: A Ten-Year Anniversary. *IEEE Transactions on Intelligent Transportation Systems*, 18(10), pp.2802-2825.
22. Pesantez-Narvaez, J., Guillen, M. and Alcañiz, M., 2019. Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression. *Risks*, 7(2), p.70.

23. Arumugam, S. and Bhargavi, R., 2020. *A Survey On Driving Behavior Analysis In Usage Based Insurance Using Big Data*.

24. Hipp,J.,2020.[online]Cs.unibo.it.Availableat: <http://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf>.

25. Sun, S. and Huang, R., 2010. An adaptive k-nearest neighbor algorithm. *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*.

26. Deng, Z., Zhu, X., Cheng, D., Zong, M. and Zhang, S., 2016. Efficient k NN classification algorithm for big data. *Neurocomputing*, 195, pp.143-148.

27. Hoffman, J., 2019. Logistic Regression. *Basic Biostatistics for Medical and Biomedical Practitioners*, pp.581-589.

28. Peng, C., Lee, K. and Ingersoll, G., 2002. An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research*, 96(1), pp.3-14.

29. Tonutti, M., Ruffaldi, E., Cattaneo, A. and Avizzano, C., 2019. Robust and subject-independent driving manoeuvre anticipation through Domain-Adversarial Recurrent Neural Networks. *Robotics and Autonomous Systems*, 115, pp.162-173.3

30. Fasnacht, L., 2018. mmappickle: Python 3 module to store memory-mapped numpy array in pickle format. *Journal of Open Source Software*, 3(26), p.651.

31. Bergstra, J. and Breuleux, O., 2020. [online] Available at: <https://www.researchgate.net/profile/Razvan_Pascanu/publication/228832149_Theano_A_CPU_and_GPU_math_compiler_in_Python/links/004635314aa30be30d000000/Theano-A-CPU-and-GPU-math-compiler-in-Python.pdf>.

32. Marwa, K., 2018. *Python Programming*. US: Tritech Digital Media.

33. McKinney, W., n.d. *Python For Data Analysis*

34. Kamalanathsharma, R.K. and Rakha, H.A., 2016. Leveraging connected vehicle technology and telematics to enhance vehicle fuel efficiency in the vicinity of signalized intersections. *Journal of Intelligent Transportation Systems*, *20*(1), pp.33-44.

35. Ngassam, R.G.N., Kamdjoug, J.R.K. and Wamba, S.F., 2018, March. Setting up a Mechanism for Predicting Automobile Customer Defection at SAHAM Insurance (Cameroon). In *World Conference on Information Systems and Technologies* (pp. 878-888). Springer, Cham.

36. Gers, F., 1999. Learning to forget: continual prediction with LSTM. *9th International Conference on Artificial Neural Networks: ICANN '99*,.

37. Sahebi, S. and Nassiri, H., 2017. Assessing Public Acceptance of Connected Vehicle Systems in a New Scheme of Usage-Based Insurance. *Transportation Research Record*, *2625*(1), pp.62-69.
38. Wang, B., Panigrahi, S., Narsude, M. and Mohanty, A., 2017. *Driver identification using vehicle telematics data* (No. 2017-01-1372). SAE Technical Paper.
39. Liotine, M., 2018. Integrating Cloud Computing with Next-Generation Telematics for Energy Sustainability in Vehicular Networks. *Mobile Computing: Technology and Applications*, p.25.
40. Neumann, T., 2017, June. Automotive and telematics transportation systems. In *2017 International Siberian Conference on Control and Communications (SIBCON)* (pp. 1-4). IEEE.
41. Taherizadeh, S., Novak, B., Komatar, M. and Grobelnik, M., 2018, July. Real-time data-intensive telematics functionalities at the extreme edge of the network: Experience with the PrEstoCloud Project. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)* (Vol. 2, pp. 522-527). IEEE.