Impact of Road Surface Conditions on Motor Insurance Risk

MSc Research Project
MSC FINTECH

Aravind Kumar Ravi
Student ID: X18102310

School of Computing
National College of Ireland

Supervisor: Mr. Victor Del Rosal

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	**Aravind Kumar Ravi**
Student ID:	**X18102310**
Programme:	**MSC FINTECH** **Year: 2020**
Module:	**Research Project**
Supervisor:	**Mr. Victor Del Rosal**
Submission Due Date:	**17th August 2020**
Project Title:	**Impact of Road Surface Conditions on Motor Insurance Risk**
Word Count:6226	**Page Count: 17 pages**

I hereby certify that the information contained in this Impact of Road Surface Conditions on Motor Insurance Risk (**My submission**) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:

Date: 17. 08. 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

Impact of Road Surface conditions on Motor Insurance Risk

Aravind Kumar Ravi
Student ID: X18102310

Abstract

Dynamic shifts in the mobility industry with solutions like ride sharing, carpooling, micro hires and, an increase in driving automation and changing demographics of drivers are posing huge challenges towards the insurance industry. The pricing of insurance products is directly tied to the risk covered by the insurers, the disruption in the mobility industry calls for a rapid shift from traditional risk profiling methodologies which hugely depend on statistical demographic risk models. This study aims to define risk from an environment centric approach. Knowledge of historic collision and their severity and using a collection of sensors should help in predicting road collisions and its severity. In order to predict this at a commendable accuracy, it is important to understand in detail the influence of features like weather road condition, and vehicle conditions influence a road collision. This study aims to study such a feature for its impact on road collisions: Road surface resistance. The work involved studying two distinct datasets from the Road Safety Authority of Ireland and Transport Infrastructure Ireland on road surface conditions for its usability in a prediction task for the actuarial application in motor insurance industry. Results and conclusions have been presented based on building a Naïve Bayes classifier with the severity of road collision as an independent variable and surface condition as the sole predictor. The critical constraints in using the features for a prediction task are also discussed.

1 Introduction

1.1 Background

The global motor insurance industry is in a transformational journey due to the increase in disruptive mobility solutions and significant advancement in driving automation technologies. Due to this Motor underwriting has become more and more challenging. This work strongly aspires to provide a significant contribution towards risk modelling for the purpose of motor insurance underwriting.

1.2 Challenges & Risks in Motor Insurance Products

The practical and characteristic constraints of the work include 1. Heterogeneity of Risk – The largest risk factor for an underwriter are road collisions, which is a spatially rare and the factors causing collision can dynamically vary with respect to each geography; 2. Type of Insurance – Insurance policies are designed to cater to different stakeholders such as vehicle manufacturer, Mobility solution provider, etc. Policies also have different inclusions, exclusions and term period. Insurance can also be designed to be traditional or parametric.

1.3 Road Collision Risk Prediction

There are several related works in the Traffic collision prediction literature, the most modern study being (yuan et al) where a Hetero-ConvLSTM approach towards prediction road collisions was proposed. The methodology involves generating a spatial grid $S = \{s_1, s_2 \ldots. s_n\}$ and a Time scale with each time period measuring 24 hours given by $T = \{t_1, t_2 \ldots. t_n\}$. The model scanned the special grid S for a given time period T to predict collisions. A feature set $F = \{f_1, f_2 \ldots\ldots f_n\}$ is used to predict the collisions, where f_x represents each independent variable used to predict a collision event.

This is a dispensable method for solving the problem: Real Time Risk Modeling for motor insurance products. However, each S_i of S is dependent on distinct feature set F to accurately predict the probability of a collision or the collision rate. Thus, selecting the feature set F for each S_i becomes important for improving the performance of the model. Model ensemble technique where a model is trained exclusively for each S_i could further improve the prediction. However, there is a wide scope of research on how to perform feature selection, especially to answer a specific business question such as: What would be the severity of a predicted collision? And what is the predicted Claim arriving from the predicted collision? - This becomes the centre of Focus of this study, as it is proven that dynamically selecting or engineering f_x for each $F(S_i)$ would yield better performance of the model which means more actionable insights.

1.4 The Motivation of the study

Motor insurance is an extremely competitive landscape with extremely low profit margins in strongly regulated jurisdictions like Ireland. Acquiring precise insights on insured profiles in real time could enable dynamic pricing and thereby acquire more low risk profiles into the underwriting portfolio. By enabling dynamic pricing based on the risk, there can be a high degree of risk mitigation performed by influencing user decisions. This reduces the overall risk underwritten by an Insurer greatly influencing the liabilities side of an insurance balance sheet. This, in turn, significantly increases the profit margins. However, in order to credibly discriminate between high risk and low risk profile, features contributing to motor insurance risk are to be extensively studied. This serves as the primary motivation of this study.

1.5 Context of the study

As described earlier in section 1.3, real time road collision risk prediction has been extensively researched upon in the road collision prediction literature. Batch processing and predicting techniques are already in industrial use. However, there are very little tangible artifacts or products that classify users based on the features contributing to a specific risk. For example, what is the size of a claim given a driver who is driving in a given strip of road at a given speed and is predicted to be in a collision? For an actionable next step, it is important to predict the severity of a collision. Which, in turn, could provide insight on the level of risk, such as claim on fatal collision, claim of serious injury, claim on vehicle damage. This could enable an underwriter to dynamically price insurance.

In order to predict the severity of a road collision, the most accessible Insurance jurisdiction is chosen: Ireland and A feature f_x that is significant to Ireland is chosen: Surface Skid Resistance. Being relatively high precipitation geography, the hypothesis is that surface skid resistance may influence road collisions significantly and reflects on the risk borne by the underwriter. Surface Resistance is studied for its usability in improving the performance of the problem. Hence, within the scope of this work, it is intended to answer the question: To what extent road surface condition & SCRIM resistance, help in predicting the severity of traffic collision in a select segment of a road network? The methodology can be further extrapolated for selecting other f_x iteratively for any given set S.

1.6 Purpose of the study

The Motor insurance industry is under a transformative journey in order to adapt to changing conditions of urban and rural mobility. With varying modes of transport and varying ownership of a vehicle, the industry is being disrupted by several key players. The risk borne by an insurer is no more straightforward and there is a strong need for planning user acquisition and risk management for the actuarial bodies of motor insurance institutions. The regulatory challenges to this problem add further complexity to this problem. Hence enabling a clear risk metric for motor insurance products, one that is scalable and generic to all the mobility platforms would be vital. This study intends to enhance the knowledge of

enabling such a risk metric by furthering the literature on road collisions from a motor insurance perspective.

1.7 Specific objectives

- To investigate the availability and usability of SCRIM feature for predicting the severity of road collisions.
- To determine the usefulness of Road Surface condition to predict the severity of Road Collision.

1.8 Research question

"To what extent road surface condition & SCRIM resistance, help in predicting the severity of traffic collision in a select segment of a road network?"

1.9 Research Hypotheses

From an insurer's perspective, risk due to road collision is a critical actuarial task. It determines the overall risk an insurer can underwrite at any given time of an insurance cycle, while also serving as a real time risk mitigation tool. Extracting insights about the severity of a predicted collision is more vital than the prediction of a collision. The claim volume and the cost overhead due to a claim are directly proportional to the severity of a road collision. Hence good predictors of severity of collisions are to be identified and evaluated to create a credible risk profiling solution that can be implemented in real time with compliance from regulatory bodies.

Hypothesis: *Road Surface condition & SCRIM resistance are a good predictor of severity of road collision.*

2 Related Work

This chapter is a discussion on the existing literature in the area of road collision prediction.

2.1 Related literature

Mohamed et al demonstrate that MARS (Multivariate Adaptive Regression Splines), which is a recently designed machine learning methodology, represents a promising approach due to its high predictive accuracy for predicting crashes at unsignalized intersections even after comparing fitted MARS and NB Models. To achieve an accurate prediction, however, it is best to pick a variable using Random Forest before fitting MARS Model Random Forest is a technique proposed by Breiman known for selecting variables from a set of variables, the R package includes the Gini "IncNodePurity" diagram for the mean decrease. NB Regression Model is also recommended as it is a valuable method for understanding traffic factors influencing safety at unsignalized intersections, such as traffic volume on major roads, the percentage of trucks on main approach, etc.

Naraya et al concentrated primarily on total crashes but this approach can be expanded to determine the severity of crashes. The study showed how important heterogeneous factors have on interstate safety in making parameter models of interchange forms. Consideration of the actions of random parameters of both segments of subpopulation to achieve heterogeneous roadway influences on the specific interchange is important. For most interchange environments, with the exception of part-diamond types, horizontal degree of curvature appears to have counter-productive effects on the frequency of the accident. Nonetheless, this effect is claimed to be more prevalent in a single point urban interchange environment, due to the sample size it needs more study to prove this. Therefore, interchange forms contribute similarly to the heterogeneity but with a constant magnitude, although with the magnitude size the heterogeneity persists in considerable form due to effects of environmental and driver behaviour.

Galal et al attempted to present accident and road safety prevention characteristics along with the estimation of accident injuries. The writers have tried to analyze and compare ANN's predictive capabilities for car accident deaths, as it is the leading cause of death in Sudan. ANN (Artificial Neuron Network) is a computer model developed for information processing and prediction, having attracted numerous engineering fields as a study of injuries with this new technique. When it can be educated on medium to large data sets, it is also believed to be a valuable method for analyzing and predicting accident casualties. Sudan has a high road accident with around 61 per cent fatality and with this new method, the model will examine many car accident incidents as a dependent variable along with the country's annual population. Many registered vehicles are taken as an independent variable and the number of paved roads. To that the accident and casualty rates, appropriate safety metrics are required in Sudan. In this country, the main cause of accidents is the negligence of the driver, bad driving and speed. To reach comparable rates, precautions need to be taken. Actions to be taken to bring accidents and casualties down would mainly include speed control, mandatory use of safety belts, control of driver behaviour, improved and marked road and intersection design, etc. Actions to be taken to bring down accidents and injuries will primarily include speed control, compulsory use of safety belts, driver conduct monitoring, improved and marked design of roads and intersections, etc.

The regression model Yajie et al NB has been proposed to resolve unattended heterogeneity issues in vehicle crash results. The discrepancy between a two-component finite mix of NB regression models with FMNB-2 (fixed weight parameters) and a varying weight parameter (GFMNB-2) on the models and the comparison of group classification from both models is studied and examined by applying them to two crash datasets. Results show that GFMNB-2 can provide more accurate classifications, reveal a more suitable source of dispersion and better statistical fit output for both road and segment crash data as well as the intersection. It is assumed that GFMNB-2 provides a better alternative to explain the complexity of the crash data and the existence of the dispersion. GFMNB-2 can be created from two distinct subpopulations, one containing sites that are vulnerable to accidents to reduce risks. According to Toronto data on GFMNB-2 analysis, it has been determined that the minor input flow is the main source of dispersion and the input flow can have a major effect on the

dispersion stage. Using minor input flow in functional form listed groups in the corresponding GFMNB-2 has the least amount of VRM however. If GFMNB-2 included both minor and major yields to a worse group classification result. Comparing length and curve density of Texas data segment are two important sources of dispersion. When segment length is used in combination with other variables GFMNB-2 will provide better performance.
Texas data used in this analysis was widely distributed, and Zou et al analyzed the data using the Sichel model, which works well with highly dispersed crash data. For this criterion, it is chosen to compare the goodness of the fit statistics of two models FBMNB-2 and Sichel model. In high dispersion caused by heterogeneity, FBMNB-2 is recommended. Sichel models should be used when widely distributed crash data are suspected, and contrasted with a finite mix of NB models for both fixed and changing weights. GFMNB-2 can be calculated using the Bayesian method, and its effect on modelling results should be investigated. NB models with different dispersion parameters may provide better statistical fit efficiency, or help explain dispersion characteristics. Small size and the low sample mean values can trigger problems with estimation; the robustness of GFMNB-2 should be examined.

Shamsunnahar et al focus on the relevance of alternative, discrete outcome frameworks to model the severity of the driver injury. Two types of models called ordered and unordered response model are compared in traffic accidents to assess the severity of the driver injury. The organized response model (MGOL) is claimed to outperform unordered response (MMNL) model. These two frameworks have, however, evoked significant debate on the implementation of the correct analytical system. The explanation for two frameworks is also affected by the underreporting problem associated with crash reports, as collateral harm and minor injury accidents can be underreported, leading to a skewed estimate. It can be argued that neither the ordered nor the unordered systems outperform each other exclusively. Although underreporting had an effect on alternative frameworks to create data sample from driver injury severity sample order, the complexity of the ordered response structure can be reduced on the single error term compared to the unordered model, which due to the multiple errors involved can lead to more complex modelling approach.

Zoi et al stated that traffic data aggression is a serious factor that causes inaccuracy in most road safety studies and this paper focuses on the incorporation of real-time data in road safety research, where separate Binomial Probit models were applied for each type of crash data over a 6-minute period and collected in the real-time moment of the accident to reduce possible bias and provide bias. Those variables include average speed for all lanes, average traffic density and lighting condition during the study of accident data involving two vehicles. During the daytime, two-vehicle rear-end crashes appear more likely than at night. Sideswipe crashes involving two vehicles were found to have a positive correlation with both the gradient per lane and route. In other words, the probability of such a sideswipe crash occurring increases on non-flat road segments and for high traffic volumes. The frequency of rear-end collisions involving two vehicles were found to be correlated with the average density of traffic and type of day. Additionally, a rear-end involving two cars is more likely to occur on Sundays and holidays, as well as high traffic density rates. Multi-vehicle collisions appear more likely to occur during daylight conditions and on flat road segments at

high speeds. On straight and flat road segments the single-vehicle accident seems more likely.

Halton used parameters and results to estimate shows that average speed was found to have fixed parameters across the road user population in low traffic conditions and traffic volume. It appears that accidents appear to increase on weekdays, but this effect varies across the population of occupants of vehicles, the type of day on which accident occurred as found to have an effect on severity outcome. In addition, lighting conditions are one cause of injury, especially during daytime. Increasing the injury was found to increase vertical road curvature. The combined effect of rainy weather and drivers experience was defined as a possible association between severity and behaviour of an inexperienced driver under weather conditions especially severe weather conditions like snow. According to the Scottish Office Central Research Unit, average speed developed under dense traffic conditions meaning higher speeds imply a significant probability for a more severe accident. Exploration of the influence of real-time traffic variables on an incident provides significant insight into the incident of occurrence. The combined effect of rainy weather and driver experience has been described as a potential association of an inexperienced driver's severity and actions under particularly extreme weather conditions such as snow.

According to the Central Research Unit of the Scottish Government, average speed established under dense traffic conditions, implying higher speeds, suggests a greater possibility for a more severe accident. Exploring the influence of real-time traffic variables on an incident provides considerable insight into the incident.

3 Research Methodology and Design Specification

3.1 Method overview

This chapter is an overview of the research method, data, techniques and tools used to test Hypotheses and arrive at an answer to the research question.

3.1.1 Method Applied

This work is a quantitative study performed on categorical features using probabilistic classifier - Naïve Bayes.

3.1.2 Probabilistic Classifier

Probabilistic classifiers are commonly used in Machine learning to perform hard classification problems typically in categorical classifications. On the contrary to a normal classifier wherein output is simply given as a function of input y=f(x), probabilistic classifiers assign probabilities to each class of an output variable for every class of an input variable, P(Y|X). The output is simply given by the class with the highest probability for a given input. It can be given by,

$$Y_{out} = \max_y P(Y= y|X)$$

3.1.3 Naïve Bayes

Naïve Bayes algorithm calculates the posterior probability of a class for a given predictor. As Naïve Bayes is a probabilistic model, the output is the class with the highest posterior probability. It typically is a calculation of probability of an event happening given an influencing event has happened. Hence, the posterior probability is the probability of an event after evidence of influencing even is observed. The most important characteristic of naïve Bayes is that it assumes every feature is independent and equally important in predicting an event. Though this is untrue in a typical everyday scenario, the algorithm performs well in classification tasks. It is seen to be used in weather predictions, spam predictions etc to name a few applications. The performance is reliable because in the case of spam prediction it is important to predict if it is spam or not. It does not really matter if the prediction is made with 51% confidence or 99% confidence.

This makes Naïve Bayes an ideal choice for this problem. The assumption that surface condition is an independent predictor of the severity of road collisions will help understand the independent contribution in the prediction task. This could be ideal for feature selection.

3.1.4 Terminologies and Formula

Bayes Theorm is given by,

$$P(A|B) = \frac{P(B$$

Where,
A is the event to be predicted
B is the event that is observed for prediction
P(A) is the priori of A, which is the probability before an evidence of B is seen
P(A|B) is the posterior probability after an evidence is seen

3.2 Tool used

Naïve Bayes is implemented using a combination of packages in Python. The process involved the application of CRISP-DM approach. The data were extracted from two important sources, 1. The Road Safety Authority of Ireland and 2. Transport Infrastructure Ireland. A Business context to the problem is first studied followed by acquiring and preparing relevant data from credible sources. Naïve Bayes Algorithm is used to perform a classification task predicting the severity of an accident. The accuracy, precision, recall & F1 Score of the resulting model serve as the evaluating parameters. The CRISP – DM methodology is selected for its effectiveness in arriving at a conclusion and flexibility offered by being able to generalize the results to other geographies or business areas. However, the

most important rationale for CRISP-DM is its familiarity in the Insurance industry allowing to reproduce the study by any independent body with well-defined feature sets.

3.3 Variables and data under the study

There are two independent sources of data that pursued this research. A Historic Collision dataset maintained by the Road Safety Authority of Ireland, and the Pavement Assessment dataset maintained and made publicly available by the Transport Infrastructure Ireland. The Historic Collision dataset is acquired by the police department in the event of a collision. Over 62 features are collected during the event of a collision including weather, surface condition, latitude & longitude of the collision, severity of the event among others. However, the features of interest to this study are the Surface Condition, Severity and the location of the collision. The Pavement assessment dataset is made available every year by the Transport Infrastructure Ireland. Pavement assessment is performed using the Sideway-Force Coefficient Routine Investigation Machine. The SCIRM machine is used in measuring the surface resistance of a strip of road. A SCRIM survey is done periodically to assess the deterioration in the quality of the road surface. The measured surface friction is given by Characteristic SCRIM Co-efficient (CSC). The CSC values typically range between 0.3 and 0.65. The SCRIM survey provides a surface co-efficient at 100m granularity. This level of detailed surface resistance knowledge is typically available across different jurisdictions in the European region with surveys done every year or every two years.

For evaluating the performance of the model with surface resistance feature as the sole predictor Accuracy, Precision, Recall and F1 Scores will be used.

4 Implementation

This chapter discusses the implementation of the proposed solution

4.1 Naïve Bayes

The Naïve Bayes model is designed to predict the severity of accident {Type} for a given input of surface condition {surface}. The training dataset is used to model for the collision data, and the output from the testing dataset is evaluated for accuracy, precision, recall and F1 score. The performance from the testing dataset is used to arrive at a conclusion on the research question and the hypothesis.

4.2 Implementation Steps

The plan of action is a six-step process as dictated by the CRISP-DM framework.

1) The first step involves acquiring in depth business knowledge. Several key stakeholders in the motor insurance industry were interviewed to gather the understanding of the present state of challenges in the industry. Interviewees included senior management from companies like Zurich Motor Insurance Ireland, Cohen & company – an insurance

accounting firm, Insurance Institute of Ireland, Insurance Ireland – independent body to represent insurance underwriters and other stake holders in Ireland and La Parisienne Assurances, an international motor insurance underwriter based out of France. It was clearly evident that risk prediction is a vital actuarial task and generating personalized insights on a given profile could be critical to mitigate risk.

2) The second step is to understand the data and perform a data quality analysis. This includes understanding the metrics and problems in the dataset. The most notable problem with the dataset is the use of Irish Transverse Mercator for geolocation of road collisions. This needs conversion to WGS84

3) Based on the understanding from step 2 the data are prepared for the purpose of modelling. Given that Naïve Bayes is used for the prediction model, the numerical features were converted into the respective categorical type. This includes changing the fundamental type of the features of interest into categorical variables which were recorded as integer.

4) Following preparation of data is modelling a predictor. For this work the collision dataset is split in an 80:20 ratio for creating the Training and Testing data sets and a Naïve Bayes model is designed to predict the severity of a road collision for a given road surface condition.

5) The evaluation process involves studying the predicted output from the testing dataset for Accuracy, Precision, Recall & F1 Score.

6) The final report is intended to portray the answer to the primary research question, its implication to business and make a case for potential future work.

5 Evaluation

These sections represent the findings & observations seen over the course of the project. It includes charts, geospatial visualizations and results as observed.

5.1 Data Sets

5.1.1 Historic Collision data of Road Safety Authority of Ireland

The historic collision dataset includes all road collisions in the Republic of Ireland between the years 2005 & 2016. The dataset included 70,469 record and 68 features. The feature of interest for this research include {ITM-x}, {ITM-y}, {surface} & {type}. ITM-x & ITM-y are the coordinates for each of the collisions. The surface feature is observed to be a categorical feature with 10 levels. However, the data description documentation includes the labels shown in Table 5.1

Table 5.1 Surface Condition Labels

Label	Surface Condition
1	Dry
2	Wet
3	Frost/Ice
4	Snow
7	Other
8	Unknown

The other levels observed include 5, 6, 11 & NULL. As no information on these levels is known, the records with these labels are dropped from the datasets. These records form up to 2% of the dataset with 1637 records. Figure 5.1 shows the spread of the surface feature.

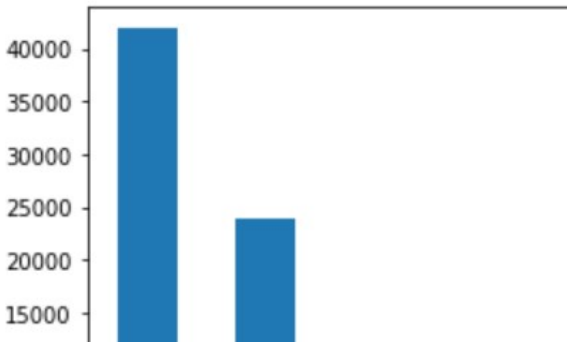

Figure 5.1 Surface Condition Distribution

The first observation from this feature is that more collisions have happened under dry conditions followed by wet conditions. The type feature has three levels, 1: Fatal, 2: Serious, 3: Minor. The spread of this feature is shown in figure 5.2.

Figure 5.2 Collision Type

The data shows significantly high records of minor collisions compared to Serious and Fatal accidents. The co-ordinates of the collisions are in Irish Transverse Mercator format. In order to perform geospatial analysis, the coordinates have been converted to WSG40 geo-coding format. Figure 5.3 visualization of the collision feature.

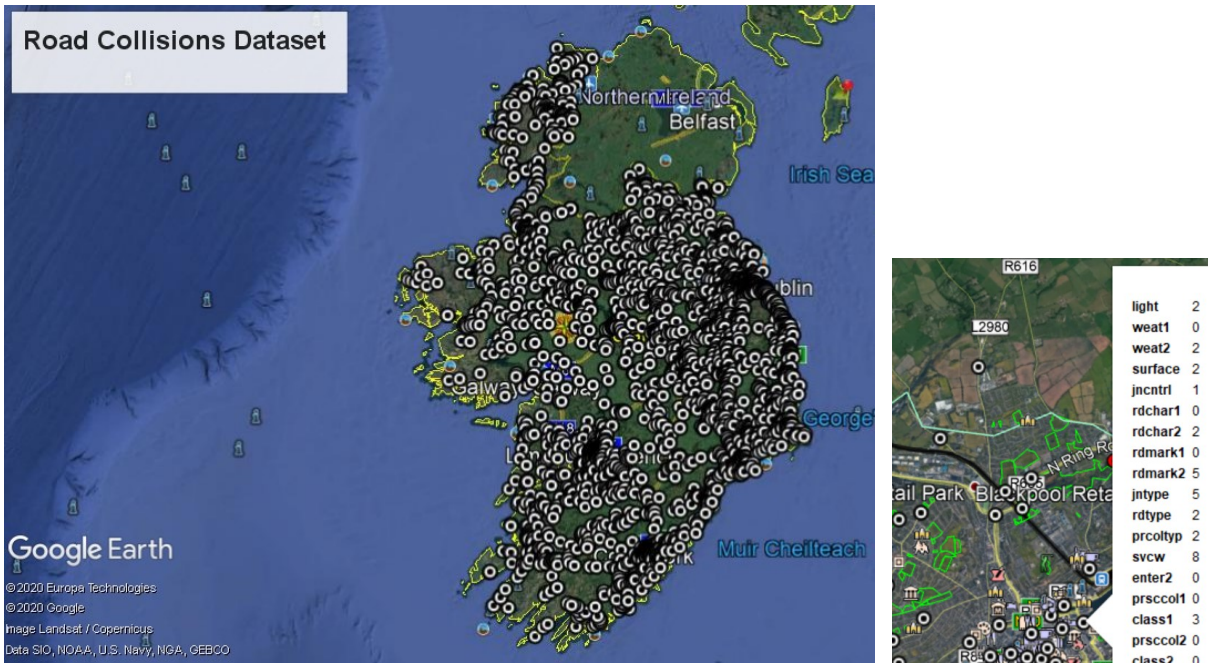

Figure 5.3 Road Collision Spread

5.1.2 Surface Resistance Dataset

The surface resistance dataset has three features of interests {csc}, {lat}, {long}. There are over 52,690 records in the dataset with a granularity of 100 meters - This can be seen in the google earth visualization shown in figure 5.4.

Figure 5.4 Surface Resistance Spread

Further analysis of the data shows that the data is available only on select roads such as National Motorways. For Example, figure 5.5 shows the same from a higher altitude.

Figure 5.5 Spread of Road Surface Resistance

This poses a huge challenge in performing any geospatial analysis or modelling for the defined scope of this project. This deems the surface resistance feature not usable. However, the surface resistance features maybe of significant use for future studies.

5.2 Testing the performance of classification

5.2.1 Experiment 1: Accuracy Score

A Naïve Bayes model is trained using the training dataset. The prediction of the model is tested using the test dataset. The Accuracy score on the prediction performed using the test dataset is measured. The test reveals 86% accuracy.

```
Accuracy score of test dataset:  0.8626425510278202
```

5.2.2 Experiment 2: Precision

Precision is the percentage of true positives over the actual results. The precision score for each class is calculated, and their weighted mean is taken as the overall precision metric of the model. The precision is measured to be 28%.

```
Precision score of test dataset:  0.2875475170092734
```

5.2.3 Experiment 3: Recall

A recall is the percentage of true positives over predicted results. The recall score for each class is calculated, and their weighted mean is taken as the overall recall of the model. It is given as below.

```
Recall score of test dataset:  0.3333333333333333
```

5.2.4 Experiment 4: F1 Score

F1 score is the harmonic mean of precision and recall. It serves as a consolidated performance metric of the classifier. Below is the F1 Score calculated from the classification made using the test dataset.

```
F1 score of test dataset:  0.30875222607859193
```

6 Results Discussion

6.1 Testing the Hypothesis

The primary hypothesis of the research is that the surface condition and SCRIM data are good predictors of severity of road collision. The result is portrayed under three categories, 1. The usefulness of a dataset/feature, 2. Performance characteristic of the model based on the dataset/feature, 3. Fitness for further modelling.

6.1.1 Usefulness of Dataset/Feature

The SCRIM resistance dataset provides extensive and precise information on the surface resistance. However, the limitation of the dataset is the data is available only for the high-speed motorways. On the average distance between the road collision dataset and a SCRIM dataset is over 2 kilometres. This means there is very little correlation with road collisions. Thus, the usefulness of this dataset/feature for the defined scope of this research is very limited. This limitation is a critical consideration for future modelling.

The Road Collision history of RSA is one other extremely useful resource. It is important to note that the data are a consolidation of all reported road collisions. This makes the dataset extremely useful from an insurance context. The severity of an incident and the road condition of each road collision is recorded among many other information. Though the surface condition is not precise like the CSC score of SCRIM dataset, the dataset has been useful for the scope of this work.

6.1.2 Performance Characteristic of the Model

The model built using the road collision dataset shows an 86% accuracy in prediction at 28% precision score and a recall score of 33.3%. The f1 score is 30.8%. This portrays that though the model performs well in classifying the false positives are higher than the acceptable rate when looking at performance per class of a given feature. This is possibly due to the huge class bias in the dataset used for the model, wherein there are significantly more minor cases of collisions compared to severe fatal cases.

6.1.3 Fitness for further Modeling

From the previous discussions, the SCRIM dataset did not serve any usefulness during the scope of the project. However, the CSC feature is a very precise and accurate measurement of road surface condition. Making the dataset critical for the purposes of future modelling. However, there is no information to answer the research question from this dataset.

The collision dataset having been solely used for the purposes of modelling provided all the information required to arrive at a conclusion. This proves the collision is the absolute minimum information required to model for risk modelling. However, the surface condition feature provides very little to the classification problem for a real time solution. This answers the question that surface feature in the collision dataset is not a good predictor of severity of road collisions in a road network.

7 Conclusion and Future Work

From the results, it is evident that surface condition is not solely a good predictor of the severity of collisions. However, having a detailed report of surface resistance could have significantly impacted the course and the result of the study. In future studies assigning a default CSC value can be given to road strips without a value. Also, a combination of CSC and satellite imaging can be used to calculate the condition of the road surface. Given that the surface condition used in this study has largely been based on weather conditions during a collision, modeling with weather conditions as an additional predictor could increase the performance of the model. This study could be used as a method to approach feature selection for the purpose of risk modeling for any insurance product.

---

8 ANNEXURE – Key Terms and Abbreviations

- **UBI:** Usage Based Insurance
- **InsurTech:** It refers to the combination of Insurance and Technology
- **Risk Profile**: Risk profile refers to the risk tied to an insured driver.
- **SCRIM:** Sideway-Force Coefficient Routine Investigation Machine
- **CSC:** Characteristic Scrim Coefficient
- **RSA:** Road Safety Authority of Ireland

9 Bibliography

1. Mohamed Abdel-Aty, Kirolos Haleem, 43(2011). Analyzing angle crashes at unsignalized intersections using machine learning techniques, pp. 461-470.

2. Narayan Venkataraman, Venky Shankar, Gudmundur F. Ulfarsson, 2(2014). A heterogeneity-in-means count model for evaluating the effects of interchange type on heterogeneous influences of interstate geometrics on crash frequencies, pp. 12-20.

3. Galal A. Ali and Awadalla tayfour, 1(2012). Characteristics and Prediction of Traffic Accident Casualties in Sudan Using Statistical Modeling and Artificial Neural Networks, pp. 305-317

4. Yajie Zou, Yunlong Zhang, Dominique Lord, 50(2013). Application of finite mixture of negative binomial regression models with varying weight parameters for vehicle crash data analysis, pp. 2042-1051.

5. Shamsunnahar Yasmin and Naveen Eluru, 59(2013). Evaluating alternate discrete outcome frameworks for modeling crash injury severity, pp. 506-521.

6. Zoi Christoforou, Simon Cohen, Matthew G. Karlaftis, 48(2012). Integrating real-time traffic data in road safety analysis, pp. 2454-2463.

7. Xin Pei, S.C Wong and N.N Sze, 43(2011). A joint-probability approach to crash prediction models, pp. 1160-1166

8. Jose M. Pardillo Mayora and Rafael Jurado Pina, 41(2009). An assessment of the skid resistance effect on traffic safety under wet-pavement conditions, pp. 881-886.

9. E. Reveron and A. Cretu. A Framework for Collision Prediction Using Historical Accident Information and Real-time Sensor Data: A Case Study for the City of Ottawa," *2019 IEEE International Symposium on Robotic and Sensors Environments (ROSE)*, Ottawa, ON, Canada, 2019, pp. 1-7, doi: 10.1109/ROSE.2019.8790431

10. Markus Deublein, Matthias Schubert, Bryan T. Adey, Jochen Kohler and Michael H. Faber, 51(2013). Prediction of road accidents: A Bayesian hierarchical approach, pp. 274-291.

11. Ni Dong, Helai Huang and Liang Zheng, 82(2015). Support vector machine in crash prediction at the level of traffic analysis zones: Assessing the spatial proximity effect, pp. 192-198.

12. Lorenzo Domenichini, Giorgio Salerno, Francesco Fanfani, Moreno Bacchi, Andrea Giaccherini, Luigi Costalli and Camilla Baroncelli, 53(2012). Travel time in case of accident prediction model, pp. 1080-1089.

13. Zhuoning Yuan, Xun Zhou and Tianbao Yang, 2018. Hetero-ConvLSTM: A Deep Learning Approach to Traffic Accident Prediction on Heterogeneous Spatio-Temporal Data

Foot note:

1. Transport Infrastructure Ireland
2. Road Safety Authority of Ireland provided the historic collision dataset dated between 2005 & 2016 on academic request.
3. Deloitte Motor Insurance Seminar 2019
4. Deloitte 13th Annual Motor Insurance Seminar