ROBERT GORDON UNIVERSITY ABERDEEN
SCHOOL OF COMPUTING SCIENCE AND DIGITAL MEDIA

INFORMATICS INSTITUTE OF TECHNOLOGY

Student Name: H L P Wickramasinghe	Matriculation Number:
Supervisor: MR. DINESH ASANKA	Second Marker:
Project Title: A Machine Learning Approach for Predicting Motor Insurance Claims	
MSc Business Analytics	Start Date: **30th September 2020**
	Submission Date: **30th April 2021**

CONSENT

☒ I agree

☐ I do not agree

That the University shall be entitled to use any results, materials or other outcomes arising from my project work for the purposes of non-commercial teaching and research, including collaboration.

DECLARATION

I confirm:

- **That the work contained in this document has been composed solely by myself and that I have not made use of any unauthorized assistance.**
- **That the work has not been accepted in any previous application for a degree.**
- **All sources of information have been specifically acknowledged and all verbatim extracts are distinguished by quotation marks.**

Student Signature: Prasad Wickramasinghe	Date Signed: 30/04/2021

Abstract

The insurance industry operates on the philosophy of uncertainty, which means the accidents and the related clams which can occur cannot be predicted. However, with the advancement of technology and statistics researchers have tried to make predictions on identifying motor claims or accidents. Integrated IoT data streams and other data sources have been used for such research. Despite much research carried out in the past, this area of study requires significant improvement and in-depth studies fill the gaps such as lack of research done at the insurance policy level using transaction data, lack of in depth knowledge in the area of study. Furthermore, such prediction capability will provide insurance companies and customers with significant financial benefits. Therefore, the research objective was to discover the possibility of identifying motor insurance policies which can have motor claims in the future and also to predict what will be the amount of the claim if a claim occurs, these predictions were done using the historical transaction data which have been recorded by insurance companies. The research looked at three main concepts which have impacts on motor insurance claims as specified by insurance experts "quality of the vehicle", "quality of the policy, and "external factors". Each concept was operationalized through a series of variables which have then mapped to the features in the database. External factors were included in the research to explore the variables that can have impact on the motor claims and claim amounts but can be external to the business. The research extracted data from a transactional database and subsequently passed on to cleansing, integrating, feature engineering, Feature Selection, training and evaluation the models. Data extraction was done using Structured Query Language where the primary data source was an Oracle database which homes historical motor insurance policy transactions. Dataset was cleansed to remove outliers and missing values. Cleansing process involved in replacing missing values, universalizing the decimal rounding in features. Python's prominent libraries numpy, pandas, sklearn along with Microsoft Excel was used to combine and cleanse the dataset. ExtraTrees algorithm was chosen to identify correlations between target variable and independent variables. During the implementation the research followed an agile approach to build and evaluate models. Models were trained using a set of predefined hyperparameters and hyper parameter values which were selected manually. Classification models were evaluated based on accuracy provided by confusion matrixes and the regression models were evaluated using Mean Absolute Error. From the results it was evident that neural network models outperformed other classification models with an accuracy of 82.53% whereas other statistical models (Random Forest, Logistic Regression, XGBoost, LightGMB) over fitted when provide full feature sets or under fitted when provided selected features. In the regression problem Random Forest Regressor was outperforming the other regression models showing the lowest Mean Absolute Error for the selected feature set. It was proven that the research supports the research hypothesis "there is a correlation between motor insurance claims and past data" or rather we are able to identify motor insurance policies which can have motor claims in the future early on before they occur. Which will have significant benefits to both end customer as well as for the insurance company as well.