Expert Systems With Applications 217 (2023) 119543

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Deep learning in insurance: Accuracy and model interpretability using TabNet

Kevin McDonnell [a,*], Finbarr Murphy [a], Barry Sheehan [a], Leandro Masello [a,b], German Castignani [b,c]

[a] KB3-040, Kemmy Business School, University of Limerick, Limerick V94 PH93, Ireland
[b] Motion-S S.A., Mondorf-les-Bains L-5610, Luxembourg
[c] University of Luxembourg, Esch-sur-Alzette L-4365, Luxembourg

ARTICLE INFO

Keywords:
Deep Learning
Telematics
Connected Vehicles
Insurance
General Linear Model
XGBoost
Machine Learning
Explainable AI

ABSTRACT

Generalized Linear Models (GLMs) and XGBoost are widely used in insurance risk pricing and claims prediction, with GLMs dominant in the insurance industry. The increasing prevalence of connected car data usage in insurance requires highly accurate and interpretable models. Deep learning (DL) models have outperformed traditional Machine Learning (ML) models in multiple domains; despite this, they are underutilized in insurance risk pricing. This study introduces an alternative DL architecture, TabNet, suitable for insurance telematics datasets and claim prediction. This approach compares the TabNet DL model against XGBoost and Logistic Regression on the task of claim prediction on a synthetic telematics dataset. TabNet outperformed these models, providing highly interpretable results and capturing the sparsity of the claims data with high accuracy. However, TabNet requires considerable running time and effort in hyperparameter tuning to achieve these results. Despite these limitations, TabNet provides better pricing models for interpretable models in insurance when compared to XGBoost and Logistic Regression models.

1. Introduction

The increasing prevalence of connected car data and advancements in Deep Learning (DL) has enhanced the ability to model driving behavior accurately. Profiling driver risk (e.g., aggressive driving behavior, context-related risk), in particular, has a societal benefit, reducing accidents and emissions. Safer driving behavior can be encouraged by using bespoke insurance products (i.e., dynamically priced insurance based on driver competency) or feedback to the driver. These bespoke insurance products such as pay-as/how-you-drive are becoming more prevalent in the motor insurance industry as insurers use a combination of telematics and Machine Learning (ML) methods for risk pricing. Traditional risk pricing models, such as the Generalized Linear Models (GLM), still dominate the insurance industry, particularly within non-life lines of business. However, in order to truly capture the relationships between the ever-expanding universe of non-traditional data (e.g. telematics, satellite, machine vision) and actuarial pricing responses (e.g., claims frequency and severity), innovative pricing mechanisms must be used. The DL method's capacity to model complex, non-linear data overcomes many of the limitations of traditional pricing models. Although, despite the increase in computational capabilities afforded by DL and the availability of highly detailed telematics data, DL is underutilized in insurance risk pricing and accident prediction. This research demonstrates the usage of an alternative DL architecture, 'TabNet', in insurance risk classification.

Deep Learning is an ML model architecture that combines or connects multiple layers to learn from data. This multi-layered approach allows each layer to learn specific traits of the presented data (Goodfellow, Bengio, & Courville, 2016). Advancements in DL have led to highly accurate models, outperforming traditional ML methods in numerous domains, such as autonomous driving, natural language processing, and marketing (Goodfellow et al., 2016). In addition, these DL models can learn complex data structures with minimal effort in preprocessing and feature engineering (LeCun, Bengio, & Hinton, 2015). However, insurance risk prediction models underutilize DL due to their 'black-box' theoretical design/framework. As a result, the explainability

* Corresponding author.
E-mail addresses: Kevin.McDonnell@ul.ie (K. McDonnell), Finbarr.Murphy@ul.ie (F. Murphy), Barry.Sheehan@ul.ie (B. Sheehan), Leandro.Masello@ul.ie (L. Masello), German.Castignani@motion-s.com (G. Castignani).

https://doi.org/10.1016/j.eswa.2023.119543
Received 26 January 2022; Received in revised form 8 July 2022; Accepted 9 January 2023
Available online 13 January 2023

or interpretability of DL insurance models is difficult to obtain (Baecke & Bocca, 2017; Paefgen, Staake, & Thiesse, 2013), which implies a challenge in terms of regulatory issues. Despite the ability of DL models to learn complex data structures, Deep Networks generalize poorly on tabular datasets compared with other Classical ML models such as Support Vector Machines, Logistic Regression and naive Bayes classifier (Arik & Pfister, 2021). Restrictions applied to granular telematics datasets cause further complications for using DL in insurance, as nascent regulation requires the anonymization and processing of raw telematics data, ensuing that tabular datasets are commonplace in insurance (Masello et al., 2022; McDonnell et al., 2021; Sheehan, Murphy, Ryan, Mullins, & Liu, 2017). Another significant limitation to their adoption in insurance is their complex overparameterized configurations (Wüthrich, 2020).

Models used for risk pricing need to be interpretable in insurance to be accepted by financial authorities (Bibal, Lognoul, de Streel, & Frénay, 2021). GLMs and Ensemble methods have proven effective in determining and quantifying risk for certain driving profiles and are regularly used in insurance risk pricing (Ayuso, Guillen, & Nielsen, 2019; Shannon, Murphy, Mullins, & Eggert, 2018; Wang & Xi, 2016; Wu, Zhang, & Dong, 2016). Additionally, GLMs are interpretable, a fundamental property required to create premium pricing models for insurance (Guelman, 2012). However, the performance of GLM can be constrained by highly complex or high dimensional data, as the linear predictor cannot accurately model high-dimensional covariate effects (Klein, Denuit, Lang, & Kneib, 2014). Ensemble methods such as Random Forest or XGBoost can learn complex data structures while also returning high levels of accuracy. However, these ensemble learning methods have drawbacks as complicated processes for both model-tuning and model interpretability can lead them to be unattractive for insurers (Pesantez-Narvaez, Guillen, & Alcañiz, 2019). For this reason, 'black-box' models are unfavorable.

The limitations of GLM, XGBoost and DL can have varying implications for an insurer's effectiveness in providing accurate pricing models for motor insurance. Additionally, both XGBoost and DL differ from GLM in their ability to provide fully explainable and interpretable pricing models, further limiting their usage. TabNet, a state-of-the-art DL architecture, addresses the limitations of DL models. Arik and Pfister (2021) introduced TabNet in their paper, citing comparative accuracy to XGBoost. TabNet utilizes single deep learning, multi-step processing, sequential attention, and gradient descent, creating an architecture that differs from traditional DL while maintaining high accuracy and providing model interpretability. The combination of these design choices makes TabNet a suitable DL model for insurance risk pricing and addresses the limitations of the other models.

To the best of our knowledge, this paper provides the first adoption of TabNet (Arik & Pfister, 2021) for insurance risk classification via claims prediction. In this article, TabNet is compared against GLM and XGBoost to demonstrate its effectiveness in accuracy, interpretability, and minimizing effort in feature training. For completeness, an additional experiment introduces a LightGBM and Neural Network to test TabNet's predictive qualities. This state-of-the-art DL architecture provides highly efficient policyholder risk prediction for insurers, using a combination of connected car data and traditional policyholder data. The implications of using this novel approach allow insurers to price individual drivers on their driving performances better when compared to traditional pricing models.

GLMs have widespread adoption in insurance risk classification and pricing due to their ability to capture the parameterized relationship between variables in driver risk classification and their response or effects (McCullagh & Nelder, 2019; Renshaw, 1994). The occurrence of claims or accidents is infrequent; therefore, distributions such as the normal distribution inadequately model these sparsely distributed accidents and claims. The GLM can vary its assumptions about the distributions of its response variables, allowing for the usage of distributions, such as Poisson regression (Ma, Zhu, Hu, & Chiu, 2018), that can better approximate relationships between risk factors and accidents. Additionally, GLMs are highly interpretable due to the generalized form of distribution and link-function selection. Due to these generalized terms, interactions between weights and variables of the GLM on a dataset can be easily interpreted (Molnar, 2020).

Combining classical policyholder data with telematics data has increased an insurer's ability to price policies accurately. Verbelen, Antonio, and Claeskens (2018) demonstrate the effectiveness of this approach by combining these traditional factors with telematics data to predict claims. The authors utilized a variant of the GLM, a Generalized Additive Model, to model risk effectively through telematics data. The GLM model highlighted exposure or mileage as a primary contributing factor to a policyholder's risk. Another variant of GLM, Poisson Regression, also provided accurate risk profile predictions for Ma et al. (2018). When combining traditional and telematics data, their GLM model identified mileage, traveling at peak times, and driving behaviors such as harsh braking as highly correlated with driver accidents. Additionally, their model identified speeding and relative speed as risk factors. Thus, GLM models can provide premium pricing models for insurers while also providing interpretable results, exposing significant features contributing to a policyholder's risk.

Although GLMs have proven effective at driver risk identification and premium pricing, there are limitations to using these models. Linear predictors find difficulty obtaining optimal solutions when learning from complex or high-dimensional data; this results in a GLM inefficiently capturing the correlations in the dataset leading to poor generalization (Klein et al., 2014). This limitation implies that models, which use linear predictors, may be unsuitable for predicting risky behaviors from high-dimensional telematics data. GLMs also require tedious manual pre-processing, feature engineering and model building processes to extract the relevant correlations and covariate effects on crash data or claims data (Henckaerts, Antonio, Clijsters, & Verbelen, 2018).

Ensemble ML algorithms have risen in popularity in the insurance domain. These machine learners can learn complex data structures, reducing the need for extensive feature engineering and provide interpretable results for insurers (Guelman, 2012). Ensemble methods combine base classifiers to produce one optimal predictive model. For example, a decision tree is a tree-based structure algorithm used to predict a class or value by learning simple decision rules. Singular decision trees, however, are prone to overfitting and variable selection bias (Quan & Valdez, 2018). Combining decision trees in an ensemble method such as Random Forest or Gradient Boosting improves the performance of these base classifiers, reducing overfitting tendencies and significantly improving the accuracy of these models.

Ensemble methods used in driver behavior risk scoring have greatly improved risk pricing models' prediction accuracy, outperforming GLMs on the same task. An important waypoint in advocating the value of ensemble methods in risk pricing is Guelman (2012), who compared GLM and Gradient Boosted Trees (GBT) on driver risk classification. The GBT model could outperform the GLM on risk pricing based on traditional risk factors and accident data. A comprehensive study by Noll, Salzmann, and Wuthrich (2018) compared variants of ensemble methods against Neural Networks and GLM. The study showed that ensemble methods performed better than GLM when predicting claims on traditional pricing features. In particular, GBT outperformed Random Forest on the same dataset. Recent studies by Pesantez-Narvaez et al. (2019) and Maillart (2021) combine the predictive power of ensemble methods with driver behavioral data. In Mailart's paper, ensemble methods were compared against a GLM, providing accurate and explainable results while reducing the need for extensive preprocessing and feature engineering.

There are limitations to using ensemble methods for insurance risk pricing. Ensemble methods have basic levels of interpretability; for an insurer, gaining sufficient levels of model interpretability for auditing or financial regulatory purposes from these models is a complex task (Henckaerts, Côté, Antonio, & Verbelen, 2021; Noll et al., 2018).

Ensemble methods can also be challenging to train. The fine-tuning of the hyper-parameters can be a difficult task and, when compared to simpler models such as Logistic Regression, the reward for the additional effort can sometimes be negligible (Pesantez-Narvaez et al., 2019).

DL has drastically changed the landscape of natural language processing, image recognition, and autonomous driving. However, these powerful models have limited usage in insurance pricing and risk classifications tasks. DL is a multi-layered approach to learning, where each layer extracts latent features and updates connected nodes and weights according to their relevance to scoring (Goodfellow et al., 2016). A mixture of feed-forward and backpropagation steps ensures that each weight is adjusted accordingly between the hidden layers of the network, where each layer outputs a vector of learned salient features, feeding this output vector to the next layer.

Studies employing DL models in insurance risk classification and pricing are limited. However, there is growing adoption of these highly accurate models. Before the growth of DL models, numerous studies tested the feasibility of DL in insurance risk pricing. Paefgen et al. (2013) demonstrated the effectiveness of DL in predicting accident risk from telematics data. The authors compare a DL model with Logistic Regression and Decision Tree models. Although the DL model outperformed the other models in various metrics, Paefgen et al. (2013) chose Logistic Regression as their favored choice due to the low interpretability of the DL model. In a similar study, Baecke and Bocca (2017) compared a DL model with a Random Forest and Logistic Regression model on driver risk classification using telematics data. Like Paefgen et al. (2013), Baecke and Bocca (2017) also decided that the Logistic Regression model was the best choice, although the DL outperformed the other models on various categories. In addition to the above studies, numerous authors introduced specially tailored DL models for insurance. For instance, in their comparative model study, (Noll et al., 2018) introduced a shallow, deep Neural Network (NN) architecture for DL on a telematics dataset for intended insurance use. (Noll et al., 2018) made significant changes to the NN model, with the highest performance achieved by introducing an exposure feature layer to the network, outperforming GLM and Ensemble methods on the task of claim prediction. An alternative DL architecture by (Siami, Naderpour, & Lu, 2021) creates a three-step approach to driver behavior extraction for subsequent risk analysis. In their paper, (Siami et al., 2021) reduce the complexity of telematics data processing the data using a self-organizing map (SOM). A 9-layer deep autoencoder extracts relevant features before a k-means algorithm clusters the dataset in the final two steps. The resulting clusters identify specific risky driver behaviors or patterns contributing to a driver's overall risk.

Learning from tabular data can be challenging for DL models to find an optimal solution, as the sparse and heterogeneous tabular datasets limit the DL model's ability to find an appropriate inductive bias (Arik & Pfister, 2021; Shavitt & Segal, 2018; Xu, Skoularidou, Cuesta-Infante, & Veeramachaneni, 2019). Additionally, due to regulation, the limitation of access to granular telematics data combined with low interpretability is a crucial obstacle to the widespread adoption of DL models in insurance risk pricing (Baecke & Bocca, 2017; McDonnell et al., 2021; Paefgen et al., 2013).

This research extends on the works of Arik and Pfister (2021), Paefgen et al. (2013), and Pesantez-Narvaez et al. (2019) by using a DL model, TabNet, for insurance risk pricing. Furthermore, this research provides the first comparison of TabNet, ML and traditional insurance pricing models. TabNet for insurance pricing offers accuracy with high model interpretability and reduced effort in data pre-processing. In addition, this model is capable of predicting accidents by combining telematics data and insurance claims. The rest of this paper is organized as follows: Section 2 describes the dataset and pre-processing steps, Section 3 outlines the methodology, Section 4 and Section 5 discusses the results in detail; finally, Section 6 provides a conclusion and future work.

2. Data

The dataset used in this study is a synthetic telematics dataset provided by So et al. (2021). This synthetic data is modeled on a real dataset provided by a Canadian insurer and generated using Synthetic Minority Oversampling Tech (SMOTE) and a feedforward Neural Network (NN) for data simulation. For evaluation purposes, the usage of Poisson and Gamma Regression ensures that the distribution of claims data links to the sparsity of actual data claims.

The dataset contains 100,000 data samples and 52 variables divided into three categories: Traditional data (Car age, Insured Age, gender), Telematic data (total miles driven, harsh acceleration, harsh braking), and Response Data (number of claims and aggregate number of claims). The breakdown of features per category is 11 Traditional Features, 39 telematics features, and 2 response variables. Table 1 contains a summary of the features and datatypes.

The response column within this dataset contains two variables, NB_Claim and AMT_claim. As per Table 1, AMT_Claim is the aggregated sum of claims paid out from the insurance company, and NB_Claim is the number of claims made by a policyholder account. A driver with NB_Claim = 1 and AMT_Claim<€1,000 may indicate first-party damage rather than a high-risk driver with third-party damages. Therefore, distinguishing between risky and non-risky drivers requires the creation of a new response column. The definition of this new response variable 'ClaimYN' is as follows: where *NB_Claim > 1 & AMT_Claim >€1,000,*

Table 1
Summary and descriptions of Traditional and Telematic variables in synthetic dataset (So et al., 2021).

Category	Feature Names	Description	Datatype
Traditional	Marital	Marital Status	Categorical
	Insured.sex	Gender: Male/Female	
	Car.use	Private/Commute/Farmer/	
	Region	Commercial	
	Territory	Rural/Urban	
		Location of Vehicle	
	Duration	Policy in days	Numerical
	Insured.age	Age in years	
	Car.age	Vehicle age in years	
	Credit.score	Credit score of Policyholder	
	Annual.miles.drive	Expected miles of driver	
	Years.noclaims	Years without claims	
Telematic	Total.miles.driven	Total Miles	Numerical
	Avgdays.week	Mean aggregated days per week	
	Accel.xxmiles*	Harsh Acc in mph (xx) per 1000 miles	
	Brake.xxmiles*	Harsh Brake in mph (xx) per 1000 miles	
	Left.turn.intensityxx*	Left turn per 1000 miles with intensity xx	
	Right.turn.intensityxx*	Right turn per 1000 miles with intensity xx	
	Annual.pct.driven	Annual percentage for time on road	Percentage
	Pct.drive.(day)		
	Pct.drive.xhrs		
	Pct.drive.wkxxx*	Driving percentage for a given day of week	
	Pct.drive.rushxx*	Hours of driving percentage for hourly period	
		Driving percentage for a given wk(end/day)	
		Driving percentage during rush hour (am/pm)	
Response	NB_Claim	Num of policy claims	Numerical
	AMT_Claim	Total claims amount	
	ClaimYN**	Risky drivers	

*xx values pre-defined buckets of values rather than random variables.
**Newly created variables, not in original dataset.

ClaimYN = 1 (driver is a risk) *else ClaimYn = 0* (driver is not a risk). Creating the response variables as per the above definition yields 97,302 ClaimYN = 0 and 2698 ClaimYN = 1 values.

3. Methodology

The following section describes the approach of classifying driver risk using driver behavioral data from telematics and comparing the performance of TabNet against GLM and XGBoost. Driver features are defined as $x_0, x_1, \cdots, x_n$, where x is a driver's feature (e.g. harsh acceleration event, harsh braking event, distance traveled) and n is the total number of features in the dataset. The target variable *ClaimYN*, for risk classification, is defined in section 2.

To demonstrate the intrinsic benefits of TabNet as a suitable risk pricing model, TabNet needs to achieve comparable or better performance than traditional risk insurance models GLM & XGBoost. This study's evaluation of TabNet compares TabNet's accuracy, model interpretability, and the reduced need for involved feature engineering processes. Comparing the accuracy of each model will use additional scoring metrics to gain further insights into each model's predictive prowess. These scoring metrics are: F1-Score, Precision, Recall, Area Under the Curve (AUC), Receiver Operating Characteristic (ROC), accuracy and Matthew's Correlation Coefficient. Evaluation of model interpretability requires each model to provide clear and insightful descriptions of data trends and identify reasonable significant risk factors. Additionally, three levels of pre-processing requirements evaluate each model's ability to provide accurate results with different levels of pre-processing effort. Models, which require less effort to provide accurate results, indicate to insurers the model's suitability for usage. Table 2 contains a summary of the evaluation methods used in this study.

3.1. TabNet

TabNet is a single deep learning model based on sequential multistep processing (Arik & Pfister, 2021). This single deep architecture assists in feature selection and improves the capacity to learn high-dimensional features. Each nth step processes a D-dimensional feature vector, where each step outputs to a Feature Transformer block. This Feature Transformer block contains multiple layers, either shared across decision steps or unique to a decision step. Each block contains fully-connected layers, a batch normalization layer, and a Gated Liner Unit (GLU) activation. Additionally, the GLU connects to a normalization residual connection; this helps stabilize the variance throughout the network. This multi-layered block assists in feature selection and increases the parameter efficiency of the network. Fig. 1 provides a detailed description of TabNet architecture.

Table 2
Summary of the evaluation methods used in this study. Model evaluation is divided into three categories, Scoring, Interpretability and Pre-processing. Scoring refers to model performance metrics such as F1-score or accuracy. Using the pre-processing methods listed, a high-scoring model with minimal user involvement indicates a good choice for an insurer. A model is highly interpretable if it can provide global and local interpretability and insights into telematics data using the interpretability methods.

Category	Method
Scoring	F1-Score
	Precision
	Recall
	Accuracy
	AUC
	ROC
Interpretability	Feature Importance
	Explain Matrix
Pre-Processing	Data Normalization
	Data Standardization
	Hyper-parameter tuning
	Feature Engineering

The Feature Transformer connects to the Attentive Transformer and Mask; these processes ensure robust feature selection per step. The Attentive Transformer is a multi-layered block with fully connected and batch normalization layers. The Attentive Transformer and masking procedure is formulated by

$$a[i-1] : M[i] = sparsemax(P[i-1].h_i([a-1])) \quad (1)$$

where $a[i-1]$ is the previous step, $P[i]$ is the priori scale and h_i some trainable function. Two key elements of the Attentive Transformer are the sparsemax activation function and the prior. Sparsemax reduces dimensionality by introducing sparsity into feature vectors; and then projecting these features onto a probability map in Euclidean space. Each projected feature vector now has an associated probability, assisting in model interpretability. The prior scale term, $P[i]$, denotes the saliency of a feature throughout the previous steps and is defined as

$$P[i] = \Pi_{j=1}^{i}(\gamma - M[j]) \quad (2)$$

where γ defines the relationship between enforcement of a feature at one decision step or multiple steps. When $\gamma = 1$ the feature is enforced at the given step and multiple steps when $\gamma = 0$. The Attentive Transformer selects the most salient features to form the transformed feature vector and passes these features to the learnable Mask, $M[j]$. The Mask enables interpretability and further improves upon feature selection from the Attentive Transformer. $M_{bj}[i]$ defines the j^{th} feature of the b^{th} sample; when $M_{bj}[i] = 0$, there is no contribution from the feature at that step. Aggregating these Masks at each step creates a coefficient that weights the importance of each step in a final decision.

TabNet can provide both local and global model interpretability, with local interpretability an intrinsic element in TabNet's design. Global interpretability is obtained through the Python library Scikit-learn (Pedregosa et al., 2011), while local interpretability is obtained by accessing TabNet's decision masks. Each mask scores features, which contributed to the model decision at that step, and each step produces a mask. TabNet is available through PyTorch version 3.1.1 (Pytorch-Tabnet, 2021).

3.2. XGBoost

Extreme Gradient Boosting (XGBoost) is an ensemble boosting tree algorithm. This fast and efficient ML algorithm can outperform other ML models in accuracy, speed, and efficiency (Chen & Guestrin, 2016). Boosting is an inherent feature of XGBoost, where previous weak learners are boosted by creating new models. These models are combined to make a final prediction. Gradient descent is used during the new model creation process to minimize the loss. XGBoost deviates from boost or gradient boosting through the incorporation of regularization (Lasso & Ridge Regression), tree creation parallelization, and tree pruning (after the tree has grown to max depth, start from the bottom and traverse up pruning invalid decisions). This approach uses feature_importances to return global model interpretability in XGBoost.

3.3. GLm

The GLM model is a generic form of the linear model. Unlike the Linear Regression model, GLM does not make assumptions on the distribution of the trainable data. For example, a normal distribution does not accurately represent sparsity in claims data. The correct choice of distribution and link-function is necessary to provide accurate predictions of risk. The GLM model can be formally defined as follows

$$g(E_Y(y \vee x)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \quad (3)$$

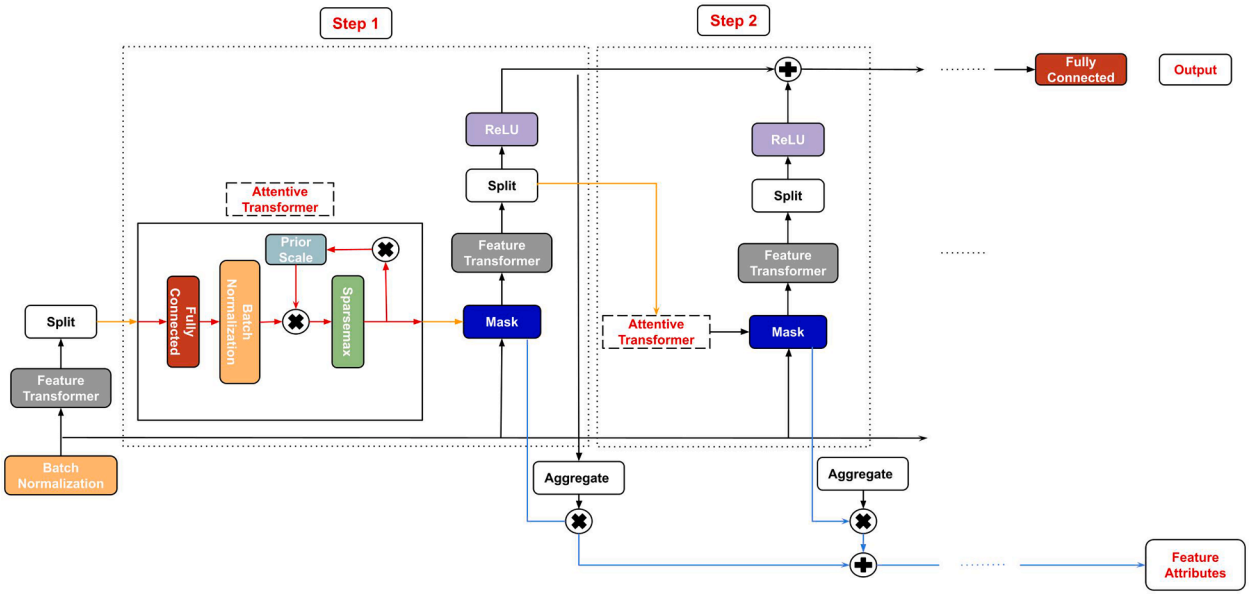

Fig. 1. Architecture of TabNet. Each step contains an attentive transformer, mask, feature transformer, split node and ReLu activation. Steps are sequential, increasing up to N steps before connecting to a fully connected layer and the output. Attentive Transformer contains a fully connected layer, batch normalization, prior scale and sparsemax dimensionality reduction. The mask function outputs significant feature contributions for aggregation. When $M_{b,j}[i] = 0$, there is no feature contribution.

where g defines the link function and E_Y the probability distribution. The choice of GLM for this study is the Logistic Regression model. Scikit-learn's implementation of Logistic Regression does not contain the feature_importances attribute; instead, the regression coefficients were used which reflect whether a feature contributed to the negative case (e.g., '$ClaimYN = 0$', if coefficient < 0); conversely, positive values indicate the positive case (e.g., '$ClaimYN = 1$', if coefficient > 0).

3.4. Pre-processing level 1–3

To demonstrate how using a DL model such as TabNet reduces the need for involved model building and feature engineering, TabNet is compared against XGBoost and GLM testing various degrees of pre-processing. The initial test restricts the capabilities of the GLM, XGBoost, and TabNet models by passing raw data to each model and running with the default hyperparameters. The only pre-processing that occurs in this step is simply encoding categorical columns. The purpose of this restricted test run is to document each model's performance with limited effort in pre-processing. Subsequent levels of testing require more involved pre-processing, feature engineering, and model building steps. The second level of pre-processing involves scaling and data normalization, along with minor model hyperparameter tuning. Data normalization for this step is simply StandardScaler and MinMaxScaler.

The final level requires more involved feature engineering and model building steps. Standardization and normalization steps continue from the previous level. The features Accel.xxmiles, Brake.xxmiles, Left.turn.intensityxx and Right.turn.intensityxx, signifies harsh driving events. These values are highly correlated and will benefit from being aggregated. Therefore, aggregating related harsh driving events to assist in this step's feature engineering requirement. The optimal hyperparameters for each model were extracted using RandomizedSearchCV with 10-fold-cross validation.

This study introduces a state-of-the-art model, LightGBM, and a Neural Network (NN) in a final test to determine the robustness of TabNet's ability to predict claims. The LightGBM and NN undergo the same training and testing processes described in 3.4. In addition, the LightGBM model requires a similar setup to the XGBoost in hyperparameter tuning. The NN, however, requires more involved tuning due to configuration requirements for hidden and dropout layers and nodes. This final test for completeness provides an additional endorsement for determining TabNet's suitability for real-life insurance uses.

4. Results

The following section describes the performance of TabNet compared with GLMs and XGBoost for both Classification and Regression tasks. For each round of model fitting and evaluation, an 80/20 split of training and test dataset combined with 10-fold cross-validation ensure the integrity of each model's performance metrics.

Table 3 illustrates each model's returned accuracy per level. Each model scores high accuracy for each level; however, due to the sparsity of driver claims, i.e., relatively few claims in the data, this metric can be

Table 3
Returned results for each model per level of preprocessing. Compared to XGBoost and Logistic Regression, TabNet scores highest in F1-Score in three out of two levels. Additionally, the model returns the highest Recall values overall. XGBoost scores highest in precision and AUC in all three rounds of testing. Logistic Regression is the worst performing model of the three. Each model returns a high accuracy value. However, high accuracy does not sufficiently represent the performance of each model in predicting claims, as demonstrated by the varying F1, Precision, Recall and AUC values. Therefore, Matthews Correlation Coefficient can assist in determining model performance on an unbalanced dataset. XGBoost and TabNet score relatively closely using this metric, although XGBoost edges TabNet in two out of three rounds.

Model Performance							
Level	Model	Precision	Recall	F1-Score	AUC	Accuracy	M Corr
1	TabNet	0.55	**0.20**	0.30	0.86	0.97	0.32
	LR	0.11	0.00	0.00	0.73	0.97	0.01
	XGB	**0.85**	0.19	**0.31**	**0.90**	**0.98**	**0.34**
2	TabNet	0.66	**0.53**	**0.59**	0.88	**0.98**	**0.58**
	LR	0.25	0.00	0.00	0.81	0.97	0.01
	XGB	**0.85**	0.37	0.51	**0.91**	**0.98**	0.54
3	TabNet	0.66	**0.55**	**0.60**	0.87	**0.98**	**0.59**
	LR	0.33	0.00	0.01	0.80	0.97	0.03
	XGB	**0.88**	0.42	0.57	**0.91**	**0.98**	**0.6**

M Corr = Matthews Correlation, LR = Logistic Regression, XGB = XGBoost.

misleading if a simple model only predicts no crash events. For this reason, f1-score, precision, recall, AUC, ROC and Matthews Correlation Coefficient are considered valuable metrics for model assessment. When comparing the three models' scores, TabNet returns the highest F1-Score for all three levels of testing, also having the highest recall scores for two out of three levels. Additionally, TabNet provides the most balanced set of results compared to the two other models. In contrast, XGBoost has returned the highest Precision and AUC score. Both TabNet and XGBoost have comparable performance with Matthews Correlation, with XGBoost scoring higher in the first round. Logistic Regression performs poorly throughout the testing, failing to predict potential claims or accidents accurately. Fig. 2 summarizes each model's performance for AUC and ROC.

As per the methodology, each round of testing requires minimal to involved preprocessing steps. For the second round of testing, XGBoost required only two changes to its hyperparameters to improve its performance, while TabNet required three, including an increase in epochs. As a result, the running time for TabNet doubled. Tuning the hyperparameters of Logistic Regression had little impact on the model performance, with a marginal increase in precision recorded for the model. Finally, RandomizedSearchCV was used to improve each model's score by selecting hyperparameters randomly in a given range, and the best score for a set of hyperparameters was retained. The optimal parameter selection had minimal impact on TabNet's and Logistic Regression's model performance, although XGBoost improved from the parameter tuning. Tuning TabNet's scores via RandomizedSearchCV was difficult due to a substantial increase in model running time, and the scope of the random search was severely limited. Despite the attempts to improve the Logistic Regression score, the model still fails to identify driver risk from the dataset. In a final test for completeness, TabNet underwent additional analysis against a LightGBM and NN. The LightGBM and NN provided disappointing results despite the increased efforts required for training. The LightGBM and NN failed to score higher than 0.2 in f1-score and Matthew's Correlation coefficient, respectively.

Model interpretability is a core component in analyzing and evaluating each model's performance in this study. Comparing each model's feature importance and model coefficients gives insight into model decisions. Table 4 contains a detailed breakdown of each model's returned feature importances. The NN and LightGBM tests did not introduce significant insights into model decisions not already captured by TabNet, XGBoost or Logistic Regression. As expected, extracting information from the NN, in particular, proved a difficult challenge.

TabNet can return the decision mask used internally for decision-making, enabling further insight into local model interpretability. For example, the first mask returned from the model shows that both traditional and telematics variables contribute to model decisions at a local level, with Right.turn.intensity11 scoring highly. Subsequent masks signify that Total.miles.driven, Accel.09miles and Duration have contributed to the model decision at a particular step. The internal mask is an inherent property of TabNet; therefore, accessing and using this property is easy. Comparing similar functionality with XGBoost and other ML models, TabNet excels at providing in-depth model analytics

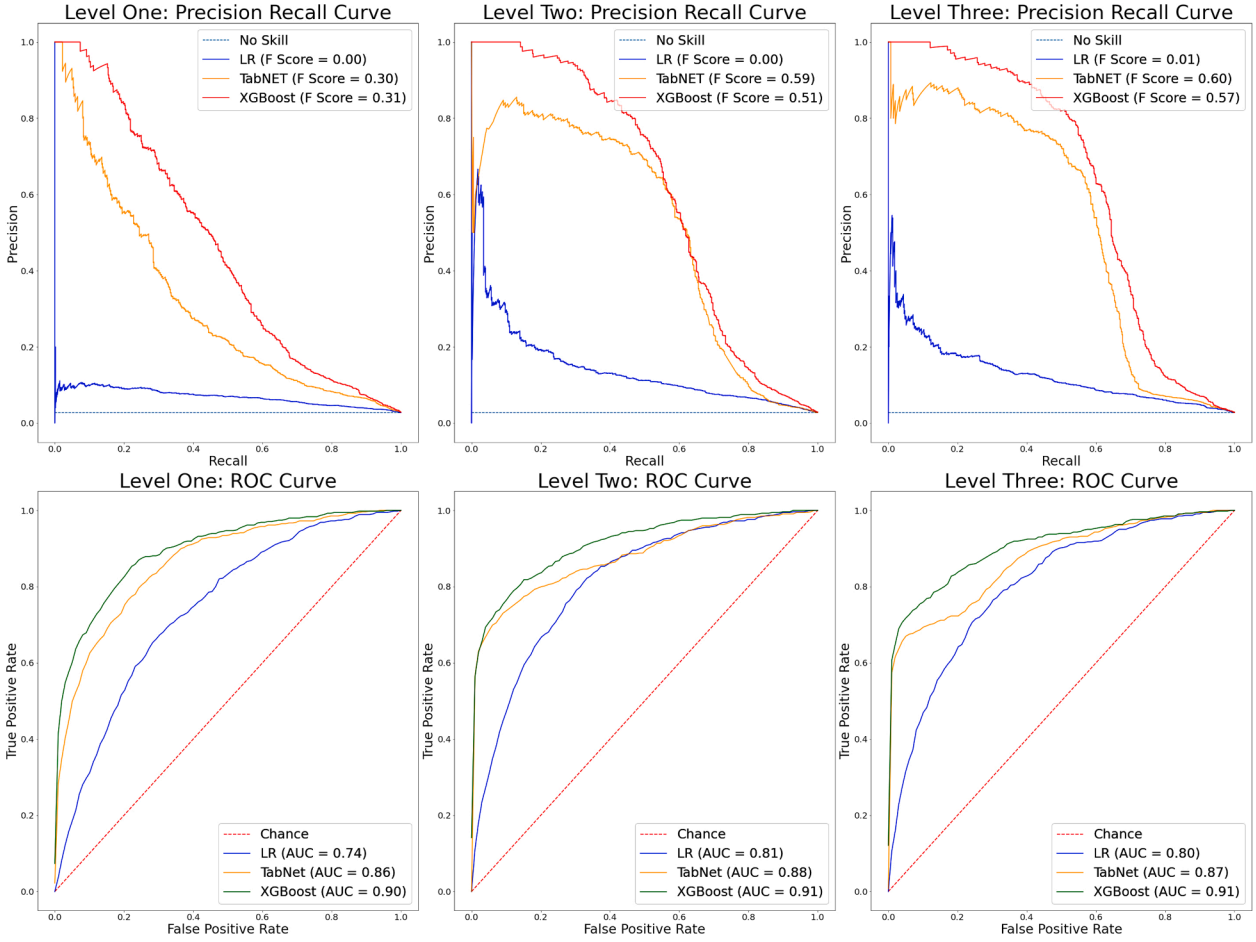

Fig. 2. AUC and ROC curves for each model per level of testing. XGBoost consistently returns the highest AUC values. However, as per the Precision-Recall Curve, TabNet returns the most balanced recall/precision values, while XGBoost scores low in recall and high in precision consistently. Additionally, TabNet scores highest in f1-score for levels two and three despite decreasing precision. Logistic Regression performs slightly better than no skill for Precision and Recall but still maintains a high AUC score.

Table 4
The top four feature importances' for each model tested per level. Bolded values are traditional or classical risk pricing variables; all other values are telematics variables. Each model used in this study has favored telematics variables over traditional or classical variables when making risk predictions. TabNet, in particular, does not use any traditional variables when making predictions. Underlined features represent commonality between each model.

Feature Importance			
Level	TabNet	XGBoost	Logistic Regression
	Feature		
1	Annual.miles.drive	***Region***	***Car.age****
	Pct.drive.wkday	Annual.miles.drive	***Duration****
	Right.turn.intensity11	Accel.12miles	Left.turn.intensity12**
	Annual.pct.driven	Accel.11miles	Annual.miles.drive**
2	Pct.drive.rush am	Annual.miles.drive	Annual.pct.driven*
	Brake.14miles	***Credit.score***	Right.turn.intensity10*
	Right.turn.intensity08	Left.turn.intensity08	Pct.drive.rush pm**
	Pct.drive.tue	Pct.drive.tue	Pct.drive.rush am**
3	Agg_Acc	***Car.age***	Annual.pct.driven*
	Annual.pct.driven	Annual.miles.drive	***Territory****
	Annual.miles.drive	Pct.drive.tue	
	Pct.drive.rush am	Brake.14miles	***Pct.drive.tue****
			Credit.score*

*Denotes coeff values that contributed to drivers identified as not a risk or 'ClaimsYN'=0.
**Denotes drivers identified as a risk or 'ClaimsYN'=1.

without additional software. In this study, only Logistic Regression's model coefficients match the ease of providing these interpretable results. Fig. 3 presents each mask per level of testing.

Table 5 contains a summation of each model's key attributes and feature interpretability capabilities.

5. Discussion

As per the results presented in Section 4 TabNet provides high levels of model interpretability and accuracy. TabNet was able to detect and classify driver risk using a combination of both traditional and telematics variables. When compared to XGBoost, TabNet can outperform this model on some classification metrics. However, XGBoost requires less intensive processing and hyperparameter tuning to achieve high levels of accuracy. When comparing each model's ability to provide interpretable results, TabNet exceeds XGBoost and GLM by returning both global and local interpretability.

5.1. Model performance

TabNet is a robust DL algorithm that has shown promising results for the task of Risk classification. However, training the TabNet model posed some challenges, which prevented the model from demonstrating the same learning capabilities as other DL models. The TabNet model tends to overfit the data and despite adjusting the available regularization parameters (the decision prediction and attention embedding layers); this had a marginal effect on the model performance. This form of regularization is not as sophisticated as DL equivalents such as dropout or weight constraint (Goodfellow et al., 2016). Additionally, criticisms of DL being over-parameterized and difficult to train also apply to TabNet. With limited hardware resources, training TabNet can take a considerable amount of time. XGBoost in comparison excels in most fields, with less extensive model building or effort required for good generalization.

The performance of Logistic Regression compared to both TabNet and XGBoost demonstrates the limitations of GLM in classifying risk on a highly dimensional dataset, including both traditional and telematics variables. Logistic Regression performed poorly on the f1-score in particular and could not provide accurate estimates of precision and recall. PCA was used to reduce the dimensionality and complexity of the dataset for the final phase of testing, although the improvement to model performance was insignificant. Logistic Regression still performed well in AUC and ROC but never surpassed the performance of TabNet or XGBoost. The poor model performance demonstrates that the model found difficulty in correctly distinguishing between risky and non-risky drivers from the telematics data.

Despite the aforementioned limitations of TabNet, the results are promising for the usage of this model in detecting claims from telematics data. TabNet could predict higher Recall and f1-scores than XGBoost and Logistic Regression on most tasks. For insurers, this is an important scoring metric, as high levels of Recall infer the model's capability in detecting true positive cases of risky driving behavior. In addition, ROC and AUC metrics do not fully represent the sparsity of claims in the dataset, high values of this score are misleading. As a result, the high f1-score and Matthew's Correlation Coefficient from TabNet signify that the model captures the sparsity in claims data, outperforming or equaling XGBoost. The addition of the final test for robustness also demonstrates TabNet's potential for usage in insurance. The LightGBM and NN models did not achieve any notable improvements in f1-score, or Matthew's Correlation Coefficient even when compared against the Logistic Regression model. In addition, the LightGBM and NN required extensive effort to achieve their training score, while TabNet and XGBoost required less training for better performance. Since TabNet is a DL model, access to larger datasets could drastically improve predictive performance. Model performance may also improve if trained on a real dataset instead of a synthetic dataset, as this data may not fully represent the chosen target variable.

The final test conducted compared an additional two models against TabNet. These models did not achieve any notable improvements in f1-score, or Matthew's Correlation Coefficient even when compared against the Logistic Regression model. In addition, the LightGBM and NN required extensive effort to achieve their training score, while TabNet and XGBoost required less training for better performance.

5.2. Model interpretability

Each model tested in this paper provided a set of features' importances or model coefficients; however, the clarity of the decisions made and the significance of these features were sometimes not as apparent. For instance, these features may be weighted differently, with some variation from the initial run. However, ranking these features and defining commonality between each returned feature for a model is beneficial, as repeated instances of the same features used may indicate a contributor to some risk cases. Once an insurer selects a production model, these features will remain constant. However, for a model to be fully interpretable, there needs to be a repeatable and easily definable method to return model decisions to provide model decision clarity for regulatory or auditing purposes (Actuarial Standards Board, 2020; Council of the European Union, 2016; Institute and Faculty of Actuaries, 2015). For example, using p-values or model coefficients, insurers or auditors can easily access model decisions from a Logistic Regression model, negating the risk of being in breach of regulation. In contrast, the features' importances returned by XGBoost score each feature uniformly before choosing the most relevant set of features. The scoring behavior of the XGBoost model is due to the ensemble decision function, where decision nodes split based on weak classifiers. These weak classifiers contribute to features' importance until a 'champion' set of features becomes part of the final solution. Additionally, the NN used in the study proved challenging to extract meaningful interior model decisions, a known disadvantage of using these models (Baecke & Bocca, 2017; Paefgen et al., 2013). The internal decision process is hidden from the user unless specialist software is used (Gramegna & Giudici, 2020). Thus, explaining model decisions and interpreting results becomes a difficult task.

Transitioning away from a GLM is difficult for insurers due to the need for model clarity and interpretability. Therefore, the black-box

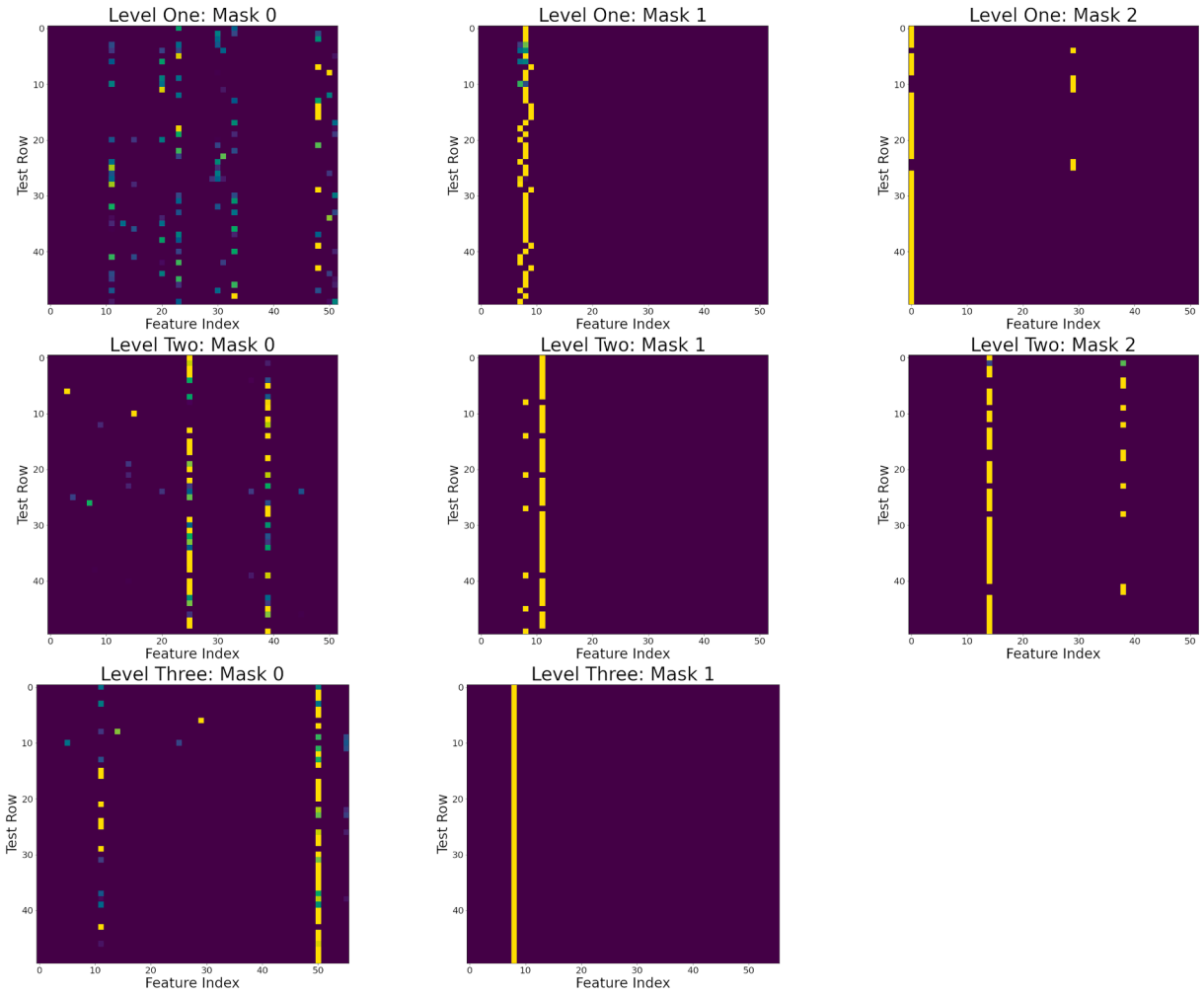

Fig. 3. Feature masks for TabNet. The highlighted features in each graph display the output from each feature mask at a given step. Each row indicates the levels of pre-processing. Additionally, color gradients signify the importance of a feature at that stage of testing, with yellow features indicating a significant contribution and purple the lowest. The x-axis refers to rows of test data, and the y-axis refers to the index of feature values. The final row has only two masks due to the optimal steps chosen as two, as each step outputs a corresponding mask.

style of model predictions found in XGB or traditional DL models becomes unattractive. However, TabNet provides both features' importance and interior decision masks. The returned feature scores from TabNet indicate a clearly chosen feature as significant to model decisions, leaving no ambiguity in model choice. Additional features that score highly are also distinct from low-scoring features. The addition of the decision mask enables further insight into the model decision process. The Attentive Transformer does not retain features that do not contribute to model decisions, and the output from the decision mask captures this process. Gaining access to these decisions is simple and does not require additional software packages. TabNet's ability to easily provide model interpretability can bridge this gap in using DL in insurance as these properties will enable insurers to provide auditors or regulatory bodies with sufficient model information to be regulatory compliant.

5.3. Data trends

The returned features' importance for each model assists in discovering trends in the telematics or policyholder data. An important consideration for insurers should be the value of investing in bespoke insurance products and whether these additional features contribute to model decisions. The returned features' importances for each model vary in their assigned importances to traditional or telematics data. These decisions from the model help insurers target certain behaviors or characteristics and offer specific and competitive bespoke insurance products. The chosen model must identify realistic and accessible features with relative accuracy to achieve this. For instance, Logistic Regression and XGBoost both identified *Car.age* and *Credit.Score* as a significant contributor to model decisions. These features are likely to infer a driver's risk; however, traditional risk metrics such as *Annual.miles.drive* have historically contributed to risk (Bian, Yang, Zhao, & Liang, 2018; Boucher et al., 2017) and thus should have been rated higher.

Telematics data offers significant insights into driver behavior and, therefore, should be an indicator of risk. However, out of the three models tested, TabNet was the only model that consistently identified telematics variables as significant contributors to driver risk. Compared to TabNet, the Logistic Regression model sporadically identified features, indicating poor model performance. The high f1 and recall scores returned by TabNet also reflect TabNet's ability to identify these risk indicators over the other models. Additionally, TabNet's ability to identify these risk indicators accurately will provide insurers with greater insights into their telematics data.

6. Conclusion

This paper provides a comprehensive overview of TabNet's

Table 5
Summation of Results. Key points summarize each model's performance throughout each round of testing. These include performance metrics, model complexity and model interpretability characteristics. For interpretability, ease of access to model decisions indicates good model interpretability.

Summary of Results		
Model	Key Points	Interpretability
TabNet	• Highest F1-Score • Recall highest in 2/3 rounds • Near equal Matthew's Correlation performance compared to XGB • Scores balanced across all tests • Longest running time • Model building is more complex than XGB for similar performance	• Identifies Telematics as significant to model decisions in all rounds • Internal model decisions are easily accessible. • Both TabNet and LR have equal levels of model interpretability
XGB	• Highest AUC and Precision • Comparable Matthew's Correlation to TabNet • Minimal effort in feature processing for good results. • Shortest runtime	• Identifies Telematics as significant to model decisions in all rounds • Poor interpretability without the using specialist software
LR	• Lowest scores in all three rounds	• Identifies Telematics variables as main contributors to risk • Coefficients easily obtained • Both TabNet and GLM have equal levels of model interpretablility

performance compared against traditional insurance risk classification methods in GLM and XGBoost. When compared against these models, TabNet surpasses the performance of XGBoost and Logistic Regression. TabNet is a highly performant DL architecture for insurers, capable of identifying driver risk from telematics data. This model also provides highly interpretable results and identifies valuable data trends, another important consideration for insurance risk pricing.

The usage of DL models in insurance risk pricing is underutilized due to their low model interpretability and poor generalization on tabular datasets. Additionally, GLM models and XGBoost require extensive model building processes or return poor model interpretability. This novel approach introduces TabNet for driver risk identification. TabNet provides high accuracy and model interpretability and can identify risky drivers from telematics and policyholder data.

Comparing the performance of each model, TabNet was best suited to capture the sparsity of claims within the dataset, returning the highest f1-scores in two out three rounds, scoring 0.59 & 0.6, respectively. Additionally, no other model scored higher than TabNet in Recall. The performance of XGBoost was also comparable to TabNet. XGBoost consistently returned high precision, accuracy, and AUC values and, in the first round of testing, returned the highest f1-score. However, Logistic Regression could not generalize on the dataset and could never accurately capture the sparsity in claims. Thus, model performance never improved despite efforts to optimize the model. Additionally, the final set of tests conducted using LightGBM and NN could not accurately predict claims. As a result, neither model could surpass TabNet's or XGBoost's scores.

This study identifies model interpretability as TabNet's key contribution for usage in insurance risk pricing. Each model could identify a significant feature that contributed to model decisions; however, XGBoost, in particular, could not provide clarity over the decisions made. On the other hand, TabNet excelled in this field, returning chosen features of significance with clarity while also providing local model interpretability. Furthermore, obtaining the decision masks from TabNet is a simple process, and each mask displays the decisions made by the model at that particular step. Finally, the coefficients returned from Logistic Regression provide interpretable results; however, clarity was difficult to obtain due to poor model performance. This problem persisted with the LightGBM and NN models, as both models performed poorly on the claims prediction task. The NN, in particular, demonstrated a significant challenge in obtaining model decisions, further endorsing TabNet's interpretable qualities.

Another important consideration for this study was evaluating each model's ability to identify telematics and policyholder data trends. Identifying specific trends in the data is invaluable to insurers, as they can offer tailored bespoke insurance products for their policyholders. Both XGBoost and Logistic Regression failed in this aspect, returning insights that would not benefit the insurer. Conversely, TabNet offered significant insights into data trends, identifying traditionally significant features such as annual mileage and telematics features as contributors to risk.

TabNet tended to overfit the data despite these aforementioned benefits and required extensive effort to train and tune the model correctly. Additionally, training the model was a time-consuming process with sometimes minimal reward. XGBoost, in comparison, could achieve comparable performance with minimal effort.

CRediT authorship contribution statement

Kevin McDonnell: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing. **Finbarr Murphy:** Funding acquisition, Supervision, Writing – review & editing. **Barry Sheehan:** Supervision, Writing – review & editing. **Leandro Masello:** Resources, Writing – review & editing. **German Castignani:** Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The dataset used in this study is a synthetic telematics dataset provided by So et al. (2021)

Acknowledgements/Funding

This project was supported by the Science Foundation Ireland (SFI), and by Lero, the SFI Research Center for Software, [grant Blended Autonomous Vehicles, BAV]. The authors would also like to acknowledge Greenval Insurance for their support in this research.

References

. Actuarial Standards Board. http://www.actuarialstandardsboard.org/wp-content/uploads/2020/01/asop056_195-1.pdf.

Arik, S.Ö., & Pfister, T. (2021). TabNet: Attentive Interpretable Tabular Learning. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(8), 6679–6687.

Ayuso, M., Guillen, M., & Nielsen, J. P. (2019). Improving automobile insurance ratemaking using telematics: Incorporating mileage and driver behaviour data. *Transportation, 46*(3), 735–752. https://doi.org/10.1007/s11116-018-9890-7

Baecke, P., & Bocca, L. (2017). The value of vehicle telematics data in insurance risk selection processes. *Decision Support Systems, 98*, 69–79. https://doi.org/10.1016/j.dss.2017.04.009

Bian, Y., Yang, C., Zhao, J. L., & Liang, L. (2018). Good drivers pay less: A study of usage-based vehicle insurance models. *Transportation Research Part A: Policy and Practice, 107*, 20–34. https://doi.org/10.1016/j.tra.2017.10.018

Bibal, A., Lognoul, M., de Streel, A., & Frénay, B. (2021). Legal requirements on explainability in machine learning. *Artificial Intelligence and Law, 29*(2), 149–169. https://doi.org/10.1007/s10506-020-09270-4

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L:2016:119:FULL&from=EN.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
Gramegna, A., & Giudici, P. (2020). Why to Buy Insurance? An Explainable Artificial Intelligence Approach. *Risks, 8*(4), 137. https://doi.org/10.3390/risks8040137
Guelman, L. (2012). Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications, 39*(3), 3659–3667. https://doi.org/10.1016/j.eswa.2011.09.058
Henckaerts, R., Antonio, K., Clijsters, M., & Verbelen, R. (2018). A data driven binning strategy for the construction of insurance tariff classes. *Scandinavian Actuarial Journal, 2018*(8), 681–705. https://doi.org/10.1080/03461238.2018.1429300
Henckaerts, R., Côté, M.-P., Antonio, K., & Verbelen, R. (2021). Boosting Insights in Insurance Tariff Plans with Tree-Based Machine Learning Methods. *North American Actuarial Journal, 25*(2), 255–285. https://doi.org/10.1080/10920277.2020.1745656
Klein, N., Denuit, M., Lang, S., & Kneib, T. (2014). Nonlife ratemaking and risk management with Bayesian generalized additive models for location, scale, and shape. *Insurance: Mathematics and Economics, 55*, 225–249. https://doi.org/10.1016/j.insmatheco.2014.02.001
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539
Ma, Y.-L., Zhu, X., Hu, X., & Chiu, Y.-C. (2018). The use of context-sensitive insurance telematics data in auto insurance rate making. *Transportation Research Part A: Policy and Practice, 113*, 243–258. https://doi.org/10.1016/j.tra.2018.04.013
Maillart, A. (2021). Toward an explainable machine learning model for claim frequency: A use case in car insurance pricing with telematics data. *European Actuarial Journal*. https://doi.org/10.1007/s13385-021-00270-5
Masello, L., Sheehan, B., Murphy, F., Castignani, G., McDonnell, K., & Ryan, C. (2022). From Traditional to Autonomous Vehicles: A Systematic Review of Data Availability. *Transportation Research Record, 2676*(4), 161–193. https://doi.org/10.1177/03611981211057532
McCullagh, P., & Nelder, J. A. (2019). In *Generalized Linear Models* (2nd ed.). Routledge. https://doi.org/10.1201/9780203753736.
McDonnell, K., Murphy, F., Sheehan, B., Masello, L., Castignani, G., & Ryan, C. (2021). Regulatory and Technical Constraints: An Overview of the Technical Possibilities and Regulatory Limitations of Vehicle Telematic Data. *Sensors, 21*(10), 3517. https://doi.org/10.3390/s21103517
Molnar, C. (2020). *Interpretable Machine Learning*.
Noll, A., Salzmann, R., & Wuthrich, M. V. (2018). Case Study: French Motor Third-Party Liability Claims. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3164764
Paefgen, J., Staake, T., & Thiesse, F. (2013). Evaluation and aggregation of pay-as-you-drive insurance rate factors: A classification analysis approach. *Decision Support Systems, 56*, 192–201. https://doi.org/10.1016/j.dss.2013.06.001
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*(85), 2825–2830.
Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression. *Risks, 7*(2), 70. https://doi.org/10.3390/risks7020070
pytorch-tabnet: PyTorch implementation of TabNet (3.1.1). (2021). [Python]. DreamQuark. https://github.com/dreamquark-ai/tabnet.
Quan, Z., & Valdez, E. A. (2018). Predictive analytics of insurance claims using multivariate decision trees. *Dependence Modeling, 6*(1), 377–407. https://doi.org/10.1515/demo-2018-0022
Renshaw, A. E. (1994). Modelling the claims process in the presence of covariates. *ASTIN Bulletin*, *24*(2), 265–285. Scopus. doi:10.2143/AST.24.2.2005070.
Shannon, D., Murphy, F., Mullins, M., & Eggert, J. (2018). Applying crash data to injury claims—An investigation of determinant factors in severe motor vehicle accidents. *Accident; Analysis and Prevention, 113*, 244–256. https://doi.org/10.1016/j.aap.2018.01.037
Shavitt, I., & Segal, E. (2018). Regularization Learning Networks: Deep Learning for Tabular Datasets. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 31). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/500e75a036dc2d7d2fec5da1b71d36cc-Paper.pdf.
Sheehan, B., Murphy, F., Ryan, C., Mullins, M., & Liu, H. Y. (2017). Semi-autonomous vehicle motor insurance: A Bayesian Network risk transfer approach. *Transportation Research Part C: Emerging Technologies, 82*, 124–137. https://doi.org/10.1016/j.trc.2017.06.015
Siami, M., Naderpour, M., & Lu, J. (2021). A Mobile Telematics Pattern Recognition Framework for Driving Behavior Extraction. *IEEE Transactions on Intelligent Transportation Systems, 22*(3), 1459–1472. https://doi.org/10.1109/TITS.2020.2971214
[dataset] So, B., Boucher, J.-P., & Valdez, E. A. (2021). Synthetic Dataset Generation of Driver Telematics. *Risks*, 9(4), 58. doi:10.3390/risks9040058.
The Regulation Board, I. and F. of A. (2015, July). *APS X2: Review of Actuarial Work*. Institute and Faculty of Actuaries. https://www.actuaries.org.uk/system/files/documents/pdf/20150122-aps-x2-final-version.pdf.
Verbelen, R., Antonio, K., & Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, *67*(5), 1275–1304. Scopus. doi:10.1111/rssc.12283.
Wang, W., & Xi, J. (2016). A rapid pattern-recognition method for driving styles using clustering-based support vector machines. *American Control Conference (ACC), 2016*, 5270–5275. https://doi.org/10.1109/ACC.2016.7526495
Wu, M., Zhang, S., & Dong, Y. (2016). A Novel Model-Based Driving Behavior Recognition System Using Motion Sensors. *Sensors, 16*(10), 1746. https://doi.org/10.3390/s16101746
Wüthrich, M. V. (2020). Bias regularization in neural network models for general insurance pricing. *European Actuarial Journal, 10*(1), 179–202. https://doi.org/10.1007/s13385-019-00215-z
Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling Tabular data using Conditional GAN. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf.