A Mining Algorithm to Improve LSTM for Predicting Customer Churn in Railway Freight Traffic

Fangcan ZHAO[1], Baotian DONG[1]*, Hongqin PAN[2], Anqi SHI[3]

[1] School of Traffic and Transportation, Beijing Jiaotong University,
No. 3 Shangyuancun, Haidian District, Beijing, 100044, China
15114244@bjtu.edu.cn, btdong@ bjtu.edu.cn (**Corresponding author*)

[2] China Railway Information Technology Center, No. 6 Fuxing Road, Haidian District,
Beijing, 100844, China
hongyawen@sinorail.com

[3] Beijing Didi Infinity Technology and Development Co., Ltd., Building B1&B2, Digital Valley,
Zhongguancun Software Park Compound 8, Dongbeiwang Road, Haidian District, Beijing, 100000, China
349356657@qq.com

Abstract: Railway freight is at risk of losing customers due to intense competition in the market. Effectively managing potential customer loss is a long-term problem for railway freight. Big data technology has been widely researched and applied in recent years, and using data mining techniques to fully extract information from railway freight ticket data and discover potential customer loss is a research topic. In this study, a mining algorithm is proposed to improve the accuracy of predicting customer churn in railway freight using Long Short-Term Memory (LSTM) network. The algorithm involves three steps: constructing the time series of railway freight volume, predicting customer churn trends using a modified LSTM model, and analyzing the characteristics of customers with different degrees of loss and the characteristics of transportation demand of lost and potential lost customers. Experimental verification was conducted using customer freight ticket data from the Shanghai Railway Bureau in China and compared the proposed modified LSTM model was compared with other commonly used machine learning algorithms for churn prediction. The results showed that the present algorithm demonstrated good accuracy and adaptability. The study proposed in this paper enriches the theoretical basis of railway freight customer management, provides an effective method for predicting customer loss in railway freight, and offers technical support for railway freight management practices.

Keywords: Rail freight, Customer churn prediction, Big Data, LSTM, Data mining, Transportation demand, Customer management.

1. Introduction

The rapid development of various transportation modes and the intensifying competition in the freight market have made it essential to deeply mine the mass freight ticket data of railway transportation to extract valuable information about customers' behavior characteristics and loss tendencies. This serves as a crucial theoretical foundation and technical support for railway freight customer relationship management, while also improving the competitiveness of the railway.

The prediction of customer churn plays a vital role in customer relationship management theory. Currently, there are several methods available to predict customer churn, such as Logistic Regression (De Caigny, Coussement & Bock, 2018), XGBoost (Chen & Guestrin, 2016), Recurrent Neural Network (Chen, He & Benesty, 2015), Multi-layer Perceptron (Kasiran et al., 2014), among others. While Recurrent Neural Network (RNN) can be used to fit sequence data, it may encounter problems with gradient explosion or gradient disappearance if the sequence is too long. On the other hand, Multi-layer Perceptron (MLP) has strong adaptive abilities but is incapable of processing time-series data. With the recent advancements in deep learning, it has been widely applied to various fields, such as image processing, natural language processing, speech processing, among others. The primary advantage of deep learning is its ability to learn advanced functions incrementally from data, reducing the reliance on domain expertise and feature extraction. Thus, the application of deep learning in the field of customer churn prediction has significantly improved prediction accuracy.

Research on customer churn in railway freight traffic is relatively limited. Zhang (2015) developed a customer segmentation model based on KGFM (Knowledge Graph Factorization Machine) and proposed a method for customer status identification using the change of customer comprehensive value. The author adopted the decision tree algorithm to construct an early-warning analysis scheme of customer churn. Zhang, Peng and Liu (2019) proposed a customer churn prediction model with a Hadoop parallel framework and C4.5 decision tree. The simulation results showed that the model

https://doi.org/10.24846/v32i2y202303 ICI Bucharest © Copyright 2012-2023. All rights reserved

achieved good accuracy and prediction ability, and the efficiency of Hadoop parallel framework improved significantly with an increase in the number of samples. Obeidat and Al-shalabi (2022) presented a novel adaptive and dynamic network routing algorithm based on a Regenerate Genetic Algorithm (RGA) with the analysis of network delays. The proposed genetic algorithm (GA) offers more efficient and dynamic solutions, even in the face of variations in network topology, network dynamics, link or node deletions, and network volume (with numerous routes). Obeidat and Yaqbeh (2022) used a method for identifying the behavior of data traffic using machine learning classifiers, including genetic algorithm to detect botnet activities. Zhang (2019) proposed a parallel SVM (Support Vector Machine) customer churn prediction method based on freight customer value classification with the KFAV model for railway bulk freight customers. For scattered white goods customers, they introduced some parameters such as social and economic attributes to establish a forecast model of customer churn, adjusting the parameter characteristics. While the above studies analyzed railway freight customer churn prediction methods from different perspectives, most of them used machine learning dichotomous classification methods and did not consider the historical time series characteristics of shippers' delivery. Yang et al. (2022) used a parallel FP-Growth mining algorithm with load balancing constraints for traffic crash data, which deepens the field of data mining. Currently, research can only identify loss, and there is no accurate identification program for identifying loss tendencies without losing the customer. However, in actual customer relationship management, customers with a loss tendency but who have not yet been lost yet have a higher current retain value. Therefore, it is of great significance to identify customers with a loss tendency, accurately and in time, to improve the service level and market share of the railway. To achieve this, a method is proposed based on the characteristics of railway freight to identify the loss of freight customers and build an effective deep learning prediction model. This method leverages information from the railway freight ticket system and the customer file system, combined with the customer shipping situation and overall railway freight market dynamics.

In Section 2, a comprehensive overview is presented, encompassing the fundamental structure, research advancements, and current applications of Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) models within the relevant field of study. Section 3 introduces an Improved LSTM Model framework specifically tailored to address the task of predicting customer churn within the context of railway freight transportation. Furthermore, in Section 4, the ticketing and customer information system of the Shanghai Railway Bureau in China is utilized. The developed customer churn prediction model is subsequently employed to estimate the likelihood of customer churn. Moving on to Section 5, representative experimental cases are carefully selected to thoroughly investigate the sequential characteristics observed among three distinct customer categories: stable customer, loss-prone customer and lost customer. Section 6 provides a concise summary of the research conducted in this paper and outlines future research directions.

2. Related Works

2.1 RNN

Recurrent Neural Networks (RNNs) have gained popularity in modeling sequential data due to their ability to capture temporal dependencies. RNNs utilize feedback connections that enable the network to use previous outputs as inputs to the current step. Soh et al. (2017) utilized RNNs in developing a personalized user interface based on user behavior. Nath, Kumbhar and Khoa (2022) applied the RNN algorithm to machine translation successfully. Yang et al (2023a; 2023b) applied real-time and fixed-time allocation to the parking space sharing model in residential areas to find a favorable RNN algorithm. The improvement of this algorithm was applied in real-time traffic accident risk. Yuan et al. (2023) considered and summarized the analysis and measurement of greenhouse gas emission from urban rail transit, and concluded that the emission algorithm should be improved. Amorim, Lopes & Silva Junior (2020) have done research on the efficiency evaluation algorithm of railway ecological simulation model. Arikan, Şen & Çam (2021) argue that GA and ABC (artificial bee colony) algorithms can optimize the energy efficiency of railways. Xue et al. (2022) applied the BP (back-propagation) neural network algorithm for calculating the railway wheel sliding to ensure the safety of railway operation.

The RNN language model consists of an input layer, a hidden layer, and an output layer. The input layer maps the text sequence to continuous word vectors, while the hidden layer captures the temporal information using the current input and the previous hidden state. The output layer estimates the probability distribution of all words in the vocabulary through Softmax. RNNs have shown promising results in a wide range of applications and are a powerful tool for modeling sequential data. It includes an input layer, a hidden layer, and an output layer, forming an acyclic neural network language model expanded by time.

2.2 LSTM

LSTM is a recurrent neural network architecture that excels in processing sequence data by effectively addressing the issues of gradient disappearance and explosion. It has strong memory ability and can learn long-term dependencies, making it suitable for tasks such as speech recognition, natural language processing, and time series prediction. Several studies have utilized LSTM, such as (Li et al., 2018) in stock return forecasting, (de Araujo, 2020) in weather simulation, (Krawiec, Junge & Hesselbach, 2021) in monthly load distribution assessment, and (Antoine, Ben Abdessalem & Frasson, 2022) in cognitive load prediction during pilot's simulated take-off. Abduljabbar, Dia and Tsai (2021) showed that BiLSTM (Bidirectional Long Short-Term Memory) performed better for variable prediction horizons for both speed and flow. Stacked and mixed Uni-LSTM and BiLSTM models were also investigated for 15-minute prediction horizons resulting in improved accuracy when using 4-layer BiLSTM networks. Recent research has focused on improving the performance of LSTM by incorporating attention mechanisms, residual connections, and convolutional layers. The modified LSTM unit can be represented mathematically by the following equations (1) - (6):

$$f_t = \sigma\left(W_{fh}h_{t-1} + W_{fx}x_t + b_f\right) \tag{1}$$

$$i_t = \sigma\left(W_{ih}h_{t-1} + W_{ix}x_t + b_i\right) \tag{2}$$

$$\tilde{c}_t = tanh\left(W_{\tilde{c}h}h_{t-1} + W_{\tilde{c}x}x_t + b_{\tilde{c}}\right) \tag{3}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t \tag{4}$$

$$o_t = \sigma\left(W_{oh}h_{t-1} + W_{ox}x_t + b_o\right) \tag{5}$$

$$h_t = o_t \cdot tanh\left(o_t\right) \tag{6}$$

The forget gate in LSTM enables the selective removal of information from the cell state. Increasing the bias of the forget gate, has been shown to improve the performance of LSTM networks by Jozefowicz, Zaremba and Sutskever (2015). Recent studies have demonstrated the efficacy of LSTM in various applications including solar radiation and photovoltaic power generation forecast (Jailani et al., 2023), water level prediction (Stefenon et al., 2023), improved optimization with LsAdam-LSTM – Reference based Long Short-Term Memory (Huang et al., 2023), data selection using R-LSTM –Refinement LSTM (Mohana, Shiva Prakash & Krinkin, 2023), and multivariable prediction with LSTM (Duan, Su & Fu, 2023). These findings highlight the versatility and potential of LSTM for various fields, and advancements such as using evolutionary algorithms and improved optimization techniques continue to improve its performance.

3. LSTM-based Prediction Model of Railway Freight Customer Churn

Based on the analysis of railway freight ticket data mining, a novel hybrid LSTM-RNN structure is proposed for the prediction model of railway freight customer churn. Specifically, the LSTM network structure is employed in this model to fully utilize the time factor and mine the deep connotation of time series for making more accurate predictions of future loss tendency. The proposed model integrates the advantages of both LSTM and RNN, which enhances the performance of the model in capturing long-term dependencies and improving the prediction accuracy. This approach represents an important step towards better understanding and improving customer retention in railway freight transportation, which is essential for enhancing the competitiveness of railway freight customer relationship management.

3.1 Improved LSTM Model

The improved LSTM model is a hybrid structure of LSTM and RNN that is used to build a forecast model for railway customer churn. The model leverages the time factor in railway freight ticket data to mine the deep connotation of time series, allowing for more accurate prediction of future loss tendencies.

The LSTM network is comprised of LSTM storage blocks, each of which consists of a layer

ICI Bucharest © Copyright 2012-2023. All rights reserved

of LSTM storage cells. Each LSTM storage block in the hidden layer has K storage units that are fully connected to K storage units at time *t-1* and time *t+1* through loop connections. The hidden layer connections are then connected to the corresponding input feature vector and output label.

The LSTM network addresses the problem of gradient disappearance that is common in RNNs by saving gradient information. The sensitivity of the nodes to the input at the first moment reveals their responsiveness, with some nodes being highly sensitive and others being completely insensitive. The input door, forgotten door, and output door are displayed on the left, lower, and upper sides of the hidden layer, respectively.

To simplify the model, all doors are either fully open or fully closed. The storage unit remembers the first input whenever it forgets that the door is open and the input door is closed. Additionally, the sensitivity of the output layer can be opened and closed through the output door without affecting the unit.

The prediction of customer loss in railway freight transportation should be viewed as a sequence labelling or classification task. This is because the shipment behavior of a customer, which is recorded on a cycle-by-cycle basis, forms the input sequence, while the loss tags allocated based on the chosen loss definition form the output sequence. Unlike traditional classification problems that assume independent and identically distributed inputs, the input sequence in a sequence classification task is dependent. If a customer engages in frequent shipping during one cycle, it is reasonable to assume that the customer will continue to do so during the following observation cycle. Conversely, if a customer exhibits little or no shipping activity during one cycle, the likelihood of shipping activity during the next few weeks is relatively low. Therefore, the time dependence of a customer's shipping behavior must be considered when modeling, and neural networks with LSTM are suitable for modeling and forecasting in this context.

3.2 Selection of LSTM Activation Function

In biological neural networks, the process of learning and forgetting is accompanied by changes in the thresholds and salience of neurons, which in turn creates new connections between neurons, allowing the brain to learn new things. In artificial neural networks such as the LSTM network, activation functions are used to simulate this process and make the network nonlinear. If the weights of inputs are simply added to the outputs of neurons, the resulting neural network would be linear. However, by applying a nonlinear activation function to the output of each neuron, the neural network can become nonlinear and approach arbitrarily complex functions. In contrast, a multilayer neural network without nonlinear activation functions would be equivalent to a single-layer neural network.

In neural networks, the activation function is used to introduce nonlinearity to the model, allowing it to learn complex relationships between input and output data. By applying a nonlinear transformation to the outputs of neurons, the neural network is no longer linear. The activation function must possess certain properties, including non-linearity, differentiability, and monotonicity. There are several popular nonlinear activation functions, including the sigmoid, tanh, and relu functions. The sigmoid function, as shown in equation (7), is a common choice for neural networks due to its smooth output that can be interpreted as a probability.

$$sigmoid\left(\chi\right) = \frac{1}{1+e^{-\chi}} \tag{7}$$

The sigmoid function is a widely-used mathematical function in various fields, including artificial neural networks, statistics, and economics. It is a non-linear function that maps input values to values between 0 and 1, which makes it particularly useful for modeling probability and binary classification problems.

One of the significant advantages of the sigmoid function is its differentiability, allowing it to be optimized using gradient-based optimization algorithms like gradient descent. This property makes it well-suited for training artificial neural networks that heavily rely on gradient-based optimization algorithms. Additionally, the sigmoid function has a well-defined and interpretable output, which makes the predictions of the model easy to understand and interpret. This property is crucial in applications such as medical diagnosis where the interpretability of the model's predictions is critical.

In the context of churn prediction, the sigmoid function is suitable for loss prediction because it can produce an output indicating the status of a customer as stable customer, loss-prone customer and lost customer. The sigmoid function can be used as the output activation function of the LSTM network, and its output can de set to threshold value to make the final prediction.

3.3 Selection of LSTM Loss Function

The effectiveness of the loss prediction model and the optimization objectives are determined by the loss function. In supervised learning, there are primarily two types of problems: classification and regression. For classification problems, samples are typically assigned to predefined categories. The cross-entropy loss function is commonly used in such problems. When there are two probability distributions, *p* and *q*, with *p* represented by *q*, the cross-entropy can be calculated as shown in equation (8):

$$H\left(p,q\right)=-\sum_{\chi}p\left(\chi\right)log\,q\left(\chi\right) \tag{8}$$

However, the output of the loss prediction model may not necessarily be a probability distribution. The probability distribution characterizes the probability law of the value of a random variable. When the total number of events is limited, the cumulative distribution function can be used to calculate the probability of any event occurring, with the sum of the probabilities of all events being 1, as shown in equation (9):

$$\forall\chi\, p\left(X=\chi\right)\in\left[0,1\right]\;\&\;\sum_{\chi}p\left(X=\chi\right)=1 \tag{9}$$

In the context of railway freight customer churn prediction, the prediction model is typically designed to classify customers into one of three categories: "stable customers", "loss-prone customers", and "lost customers". In this case, the output of the model needs to be transformed into a probability distribution over the three classes. This can be achieved using the Softmax function, which compresses the k-dimensional vector of any real number into another k-real vector, so that the value of each element is within the range 0 to 1, and the sum of all elements equals 1.

To define the probability values for the three classes, a common approach is to use one-hot encoding. This means that, for each customer, the target output vector is a 3-dimensional vector where the element corresponding to the true class is set to 1, and all other elements are set to 0. For example, if a customer belongs to the "lost customers" category, the target output vector would be [0, 0, 1]. By training the model to predict the correct target output vector for each customer, the Softmax function can then be used to transform the model's output into a probability distribution over the three classes.

In the context of customer churn prediction, the Softmax function can be applied to the output of a model to obtain a probability distribution over the three possible classes: stable customers, loss-prone customers, and lost customers.

Suppose that the output of the forecast model without Softmax regression is $y_1, y_2, y_3, \ldots, y_n$. The output after Softmax regression is shown in equation (10):

$$soft\,max\left(y_i\right)=y_i^{'}=\frac{e^{y_i}}{\sum_{j-1}^{n}e^{y_j}} \tag{10}$$

Here, y_i' is the original output of the model for class *i*.

The output after Softmax regression represents the probability that a sample belongs to each of the three classes. The probability distribution obtained can be used to calculate the distance between the predicted probability distribution and the true probability distribution of customer churn, which is done by using the cross-entropy loss function.

The cross-entropy loss function measures the dissimilarity between two probability distributions *p* and *q*. In the context of customer churn prediction, *p* represents the true probability distribution over the three classes, and *q* represents the predicted probability distribution. Since the goal is to predict the true probability distribution, *p* is used as the reference distribution and *q* is the predicted distribution. The smaller the distance between the two probability distributions, the better the prediction accuracy.

The cross-entropy loss function is not symmetric, meaning that the order of *p* and *q* matters. Specifically, the formula for cross-entropy loss is the one displayed by equation (11):

$$L\left(p,q\right)=-sum\left(pi*log\left(qi\right)\right) \tag{11}$$

ICI Bucharest © Copyright 2012-2023. All rights reserved

where i is the index over the classes, pi is the true probability of class i, and qi is the predicted probability of class i.

In TensorFlow, the cross-entropy loss function is often used together with the Softmax function, and a single function is provided that encapsulates the two functions. The cross-entropy loss function after Softmax regression can be calculated directly using the following equation (12), where y represents the output from the original prediction model and y_ is the true probability distribution over the three classes:

$$\begin{aligned} &Cross\_entropy = \\ &tf.nn.softmax\_cross\_entropy\_with \\ &\_logits\left(y, y\_\right) \end{aligned} \tag{12}$$

Overall, the Softmax function and cross-entropy loss function are powerful tools for customer churn prediction. By transforming the output of a model into a probability distribution over the three classes, Softmax regression enables the use of the cross-entropy loss function to train the model and improve its prediction accuracy.

3.4 Research on Algorithm of Preventing Over-fitting

To address the problem of overfitting, a method to improve the generalization ability of the LSTM model, when predicting customer churn, is proposed. The early stopping strategy is used to prevent overfitting during the training process. Specifically, the accuracy rate of the validation set is monitored during the training process. If the accuracy rate does not improve for ten successive epochs, the training is stopped in advanced, in order to limit the ability of the model and prevent overfitting.

To evaluate the performance of the trained model, a test set consisting of samples that are not included in the training and validation sets is used. The test data is input into the trained model and the results of the test are evaluated. The evaluation results show that the proposed LSTM model achieves high accuracy and outperforms other traditional machine learning models in predicting customer churn.

The customers are classified into three categories: stable customers, loss-prone customers, and lost customers. The proposed LSTM model can accurately predict which category a customer belongs to, based on his historical behavior data.

In conclusion, the proposed LSTM model with early stopping strategy can effectively prevent overfitting and improve the generalization ability of the model. The model achieves high accuracy in predicting customer churn for the three classification categories, which provides a valuable reference for businesses to retain their customers and improve customer satisfaction.

3.5 Summary of the Improved LSTM Model Training Process for Customer Churn Prediction

The improved LSTM model for customer churn prediction involves several steps and the specific training process of the proposed model can be summarized as follows:

Step 1: Data Entry. To begin with, the model parameters are initialized and the pre-processed data set is input to the model. Data pre-processing is an essential step to ensure that the input data is clean, consistent, and structured. This step includes data cleaning, feature selection, and normalization;

Step 2: Forward Propagation. The LSTM network model is used to perform forward propagation calculations, from input features to output probabilities. In this step, the model architecture includes multiple LSTM layers and a Softmax output layer to predict the probability of each class;

Step 3: Backward Propagation. In the next step, the error term of each neuron is calculated along the time direction and the space direction, and is used to update the weights of the LSTM layers. Backpropagation through time (BPTT) is employed to calculate the error term;

Step 4: Gradient Calculation. The gradient of each neuron is calculated using the error term and backpropagation. The gradient is used to determine the direction and magnitude of weight updates;

Step 5: Weight Update. The weights of the LSTM layers are updated using the calculated gradient and learning rate. The optimization algorithm used for weight updates can be stochastic gradient descent (SGD) or its variants such as Adam or RMSprop;

Step 6: Training Termination. The training process is stopped when the validation loss has not improved for a certain number of epochs, or when the maximum number of epochs has been reached. This helps prevent overfitting and ensures that the model has learned generalizable patterns in the data;

Step 7: Result Output. Finally, the trained model is evaluated using the test data set, and the predicted classes and probabilities for each sample are output. Evaluation metrics such as accuracy, precision, recall, and F1 score can be used to evaluate the model performance.

In conclusion, the proposed improved LSTM model for customer churn prediction can effectively capture temporal dependencies in the input data and predict the likelihood of customers churning. The optimized model architecture and training process result in high accuracy and robustness, which can help businesses make data-driven decisions to improve customer retention.

4. Analysis of Individual Prediction of Railway Freight Customer Churn

Based on the ticket and customer information systems of the Shanghai Railway Bureau in China, this study estimated customer churn using the customer churn prediction model outlined in Section 3. The railway freight ticket data was utilized to track the delivery behavior of railway freight customers and further explore the dynamic characteristics of customer churn trends. This was combined with real-time data updates to enable timely and accurate understanding of railway freight customer loyalty and loss trends.

4.1 Experimental Data

The data for this study was sourced from the ticket system and customer information system of the Shanghai Railway Bureau, containing cargo ticket information data for two categories of goods, namely smoking and metal, collected between January 1, 2015 and December 31, 2018. A total of 1,620,935 railway freight ticket information data of 1,294 customers over four years was used in the model.

To construct the experimental dataset, the customer's ticket data from January 1, 2015 to December 31, 2017 was used as the input variable and also as an influencing variable, in case of any freight behavior from January 1, 2018 to December 31, 2018. The customers were then randomly divided into training (60%), validation (20%), and test (20%) sets. Cross-validation was also performed to further evaluate the performance of the model.

4.2 Rail Freight Ticket Data Pre-processing

Data pre-processing is a crucial stage in data mining. An effective data pre-processing scheme can significantly enhance data quality and decrease computation costs, thereby enabling the prompt generation of high-quality analysis outcomes.

4.2.1 Data Cleaning

Railway freight ticket data constitutes a large-scale data collection with high dimensionality. In this regard, a distributed storage and a parallel IForest cleaning method were proposed for railway freight ticket data, as illustrated in Figure 1.

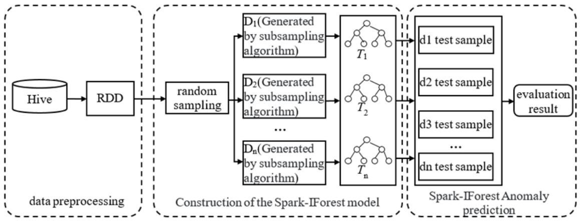

Figure 1. The process of Spark-IForest parallel algorithm data cleaning

IForest, short for Isolation Forest, is a tree-based anomaly detection algorithm that constructs a random tree to judge whether the data is an outlier. Traditional anomaly detection algorithms become unreliable for high-dimensional data, such as railway freight ticket data, as the distance measurement increases with the increase of data dimensions. In contrast, IForest uses subspace division to calculate the average height of data objects in the IForest model in order to achieve anomaly instance detection. The Spark-IForest parallel algorithm has linear time complexity and good scalability for high-dimensional data, making it suitable for processing railway freight ticket data.

ICI Bucharest © Copyright 2012-2023. All rights reserved

4.2.2 Fusion of Multidimensional Features for Railway Freight Customer Integration

Based on the dataset of railway freight customer ticket information, the comprehensive behavioral sequences of railway freight customers were extracted. By analyzing the overall distribution characteristics and individual-level temporal features, the sequential properties of railway freight customer behavior are comprehensively explored. This study provides a theoretical foundation and data support for constructing a customer churn prediction model.

Railway freight ticket data is a multidimensional and ordered dataset that captures information about delivery, mileage, cargo, and billing for each transportation process of freight customers. Focusing on railway freight customers, multidimensional time series data are extracted from six clusters of information: volume, category, station, railway bureau, mileage, and cost. These dimensions collectively construct the multidimensional comprehensive behavioral sequence data of freight customers within a specific observation time window.

The extraction of comprehensive behavioral sequences from customers is performed by processing the cleaned ticket data through the following steps:

Step 1: Extract all consignor IDs from the railway freight ticket data;

Step 2: Aggregate the data based on consignor IDs to obtain all ticket information for each customer;

Step 3: Sort the ticket information for each customer based on the ticket generation time to create the customer's ticket information time series data;

Step 4: From the ticket time series data for each customer, various features can be extracted, including the dimension of volume, category, station, railway bureau, mileage, and cost. These features are aggregated based on the shipping dates.

Taking consignor ID 3190777 from experimental data as an example, partial segments of the comprehensive behavioral sequence can be observed in Figure 2.

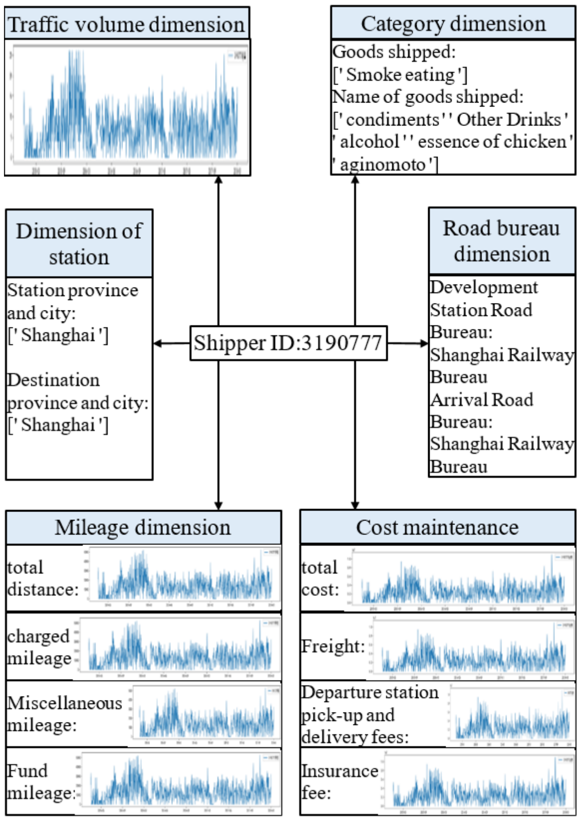

Figure 2. Behavioral sequence representation for ID 3190777

The railway freight customer behavior sequence exhibits high dimensionality, density, and nonlinear dynamic changes. Different customers display significant variations. The internal multidimensional collinearity of the sequence manifests periodicity, trends, and volatility.

To predict customer churn in railway freight transportation, a feature set based on the comprehensive customer sequence is constructed.

4.2.3 Dynamic Sequence Features

Due to the strong collinearity among the dimensions of freight, mileage, and traffic volume, the comprehensive behavior sequence of railway freight customers, incorporating the railway freight volume sequence as a dynamic feature, makes it possible to capture the transportation demand and activity of customers during different time periods.

4.2.4 Static Features

The static characteristics of customer behavior, such as the "destination" and "transportation category", are extracted. The static features, derived from the customer's departure and arrival information, provide valuable geographical location insights into their transportation needs. The transportation category information of customers may span across various industries and markets.

By integrating multiple dimensions of features, including dynamic features such as the railway freight volume sequence, as well as static features such as the customer's departure and arrival information and transportation category information, the customer churn can be accurately predicted by comprehensively analyzing the temporal and transportation demand characteristics, and geographical location information of customer behavior sequences. Additionally, this approach allows for a detailed description of customer characteristics and provides a solid data foundation for churn prediction.

4.3 Feature Engineering

(1) Dynamic Sequence Features Engineering

The volume series of railway freight transport customers exhibit characteristics about trend, cycle, and fluctuation. As the loss of individual railway freight customers primarily depends on the gradual decline of the customer's shipment trend, the trend sequence of the freight volume sequence is extracted and the non-trend changes, such as periodicity and volatility, are eliminated. To achieve this, the Prophet model is utilized to extract the trend sequence of railway freight customers. The Prophet model decomposes the time series into four types of items: trend model, periodic (seasonality), holiday model and random model, as shown in equation (13):

$$y(t) = g(t) + s(t) + h(t) + e(t) \tag{13}$$

Here, $y(t)$ represents the time series value at time t, g(t) is the trend model simulating the non-periodic change of the time series value, $s(t)$ is the periodic model indicating the periodic influence effect of the time series, $h(t)$ is a holiday model representing changes caused by special circumstances such as holidays, and $e(t)$ is a random model representing unpredictable fluctuations.

The trend model fits the non-periodic changes of the sequence and specifies the change point that affects the time series trend, adaptively adjusting the position of each change point and the linear trend, as seen in equation (14):

$$g(t) = \left(k + a(t)^T \delta\right)t + \left(m + a(t)^T \gamma\right) \tag{14}$$

Here, k represents the growth rate, $\mathbf{a}(t)^{\mathrm{T}}$ is a binary vector indicating whether the time t is a change point, δ represents the change in growth rate, m is the offset parameter, and γ is the parameter making the trend model continuous.

(2) Static Feature Engineering

One-hot encoding is utilized for feature engineering of static features. One-hot encoding enables the transformation of discrete categorical features into a numerical format, offering an effective means to represent and process these features while retaining their information and interpretability.

To convert the "origin-destination" and "transportation category" features of customers into one-hot encoding, the next process is followed:

1. "Origin-Destination" Feature

All unique combinations of origin and destination are identified in the dataset. Each unique combination is assigned a binary column in the feature matrix. For each customer data point, the corresponding binary column is set to 1 to indicate the specific origin-destination combination, while all other columns are set to 0.

2. "Transportation Category" Feature

All distinct transportation categories are identified in the dataset. Each transportation category is assigned a binary column in the feature matrix. It is similar to the "origin-destination" feature, for each customer data point, the corresponding binary column is set to 1 to represent the specific transportation category, and the rest of the columns are set to 0.

3. Integration into Feature Vector

The one-hot encoded columns of both features (origin-destination and transportation category) are concatenated to form a unified feature vector.

This feature vector represents the customer's static features, with each element in the vector indicating the presence or absence of a specific origin-destination or transportation category.

(3) Concatenation of Dynamic and Static Features

The process of concatenating the dynamic and static feature vectors involves merging

ICI Bucharest © Copyright 2012-2023. All rights reserved

the two feature vectors into a single, unified representation. This concatenation operation combines the information from both feature components to create a comprehensive feature vector that captures both the temporal dynamics and the relevant categorical information of the customer's transportation behavior.

To perform the concatenation, the dynamic feature vector, representing the customer's transportation volume time series, and the static feature vector, encoding the "origin-destination" and "transportation category" attributes, are aligned and appended together. This operation ensures that the corresponding feature values of each customer are correctly preserved in the combined feature vector.

By concatenating the two feature vectors, a new feature vector that contains both the time-dependent patterns of the customer's transportation volume and the categorical information related to their transport characteristics is created.

This unified feature vector serves as a comprehensive representation of the customer's transportation behavior, enabling efficient utilization of the combined information for customer churn tasks.

4.4 Evaluation Metrics

In order to evaluate the performance of the churn prediction models, the commonly used evaluation metrics are applied, including accuracy, precision, recall, and F1-score. These metrics were calculated based on the fundamental components of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These indicators provide essential information for assessing the performance of a classification model. TP refers to correctly identifying positive instances, TN refers to correctly identifying negative instances, FP refers to incorrect positive predictions, and FN refers to incorrect negative predictions.

1) Accuracy: The proportion of correct predictions among all predictions made, as shown in equation (15):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{15}$$

2) Precision: The proportion of true positive predictions among all positive predictions made, as shown in equation (16):

$$Precision = \frac{TP}{TP+FP} \tag{16}$$

3) Recall: The proportion of true positive predictions among all actual positive instances, as shown in equation (17):

$$Recall = \frac{TP}{TP+FN} \tag{17}$$

4) F1-Score: The harmonic mean of precision and recall, providing a single score that balances both metrics, as shown in equation (18):

$$F1 = Score = 2 * \frac{precision * recall}{precision + recall} \tag{18}$$

In summary, accuracy, precision, recall, and F1-score provide a comprehensive evaluation of the performance of the railway freight customer churn prediction model in identifying customers who are likely to churn.

4.5 Hyperparameter Tuning for LSTM Model Optimization in Churn Prediction

LSTM is a type of neural network with several hyperparameters that can be tuned to optimize the performance of the model during training.

To train an LSTM model for churn prediction using customer traffic sequence as input and customer classification as output, it is necessary to tune the hyperparameters of the LSTM model. The following hyperparameters and their corresponding values were selected based on experimentation and evaluation:

Number of LSTM layers: 2

Number of LSTM units: 64

Dropout rate: 0.2

Learning rate: 0.001

Batch size: 32

Number of epochs: 50

Optimization algorithm: Adam

These results indicate that the LSTM model, with the selected hyperparameters, can accurately classify customers as "stable customers", "customers with churn tendency", and "customers who have churned", based on their traffic sequence data.

LSTM hyperparameter tuning involves adjusting the architectural and training hyperparameters to find the optimal combination of hyperparameters that maximizes the performance of the model.

4.6 Comparison of Churn Prediction Model Performance

To demonstrate the effectiveness of using an improved LSTM for churn prediction, the performance of the proposed model was compared with the performances of other popular machine learning models commonly used for churn prediction. Table 1 presents the evaluation results of various methods, including the improved LSTM, for predicting customer churn.

Table 1. Comparison of churn prediction model performance

Model	Accuracy	Precision	Recall	F1-score
Improved LSTM	0.85	0.82	0.87	0.84
Random Forest	0.81	0.79	0.81	0.79
LR	0.76	0.72	0.76	0.74
SVM	0.80	0.78	0.80	0.78
XGBoost	0.82	0.80	0.82	0.80
LSTM	0.81	0.79	0.81	0.79

The evaluation results indicate that the improved LSTM method outperforms the other methods evaluated, including Random Forest, LR, SVM, XGBoost, and the traditional LSTM model. The improved LSTM model demonstrated higher accuracy, precision, recall, and F1-score, suggesting that the modifications made to the LSTM architecture led to an improved performance of the model.

Based on these findings, the present paper recommends the use of the improved LSTM model for customer churn prediction.

5. Analysis of Sequence Characteristics of Three Customer Categories

This section focuses on extracting typical experimental cases to analyze the sequence characteristics of three distinct customer categories: stable customer, loss-prone customer and lost customer. This analysis aims to contribute to a better understanding of customer behavior and facilitate the development of more effective churn prediction models.

5.1 Characteristics of Stable Rail Freight Customers

Stable customers are defined as those who exhibit a consistent and predictable freight volume over time, with minimal fluctuations or irregularities in their shipment behavior. This behavior is represented in their time series data, which shows a relatively flat trend, with only minor fluctuations and no clear signs of decreasing volume over time.

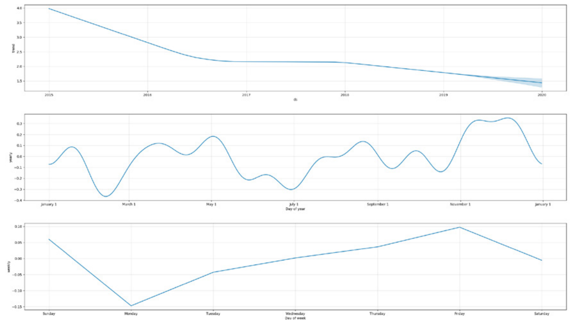

Figure 3. Volume time series for railway freight customer code 3001223

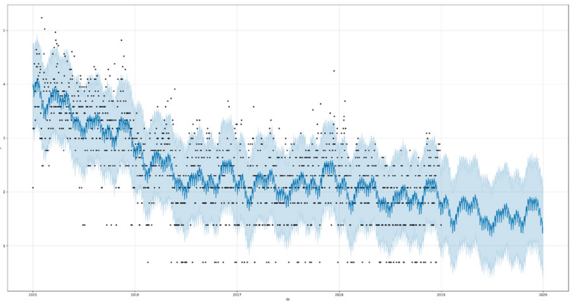

Figure 4. Traffic forecast for railway freight customer code 3001223

Using the improved LSTM algorithm, the customer with shipper code 3001223 was classified as a stable customer. The time series information of railway freight customers between 2015 and 2017 was extracted, and trend items were extracted accordingly, as illustrated in Figure 3. The user's future traffic was predicted, as depicted in Figure 4. It can be observed that the trend item in the time series of transportation behavior is relatively stable and flat, indicating consistent and regular shipment patterns.

Their time series data shows a relatively flat trend, with only minor fluctuations and no clear signs of decreasing volume over time. These customers are considered reliable and valuable to railway freight companies, as they provide a stable and predictable source of revenue. In contrast to loss-prone and lost customers, stable customers do not exhibit significant changes in their shipment

ICI Bucharest © Copyright 2012-2023. All rights reserved

behavior, making them less of a priority for churn prediction efforts.

5.2 Characteristics of Rail Freight Customers with a Tendency to Churn

Rail freight customers who exhibit a tendency to churn are characterized by a gradual decrease in shipment volume over time, which indicates a loss of interest in the rail freight service, resulting in decreased demand and ultimately leading to their departure from the service. These customers often exhibit erratic and inconsistent shipment patterns, with irregular order placements and an overall reduction in the frequency of shipments.

Taking shipper code 3196623 as an example, the present analysis reveals that this shipper exhibits a significant tendency towards churn. The time series information of railway freight customers between 2015 and 2017 was extracted and depicted in Figure 5, and the trend items were extracted accordingly. The future traffic volume of the user was predicted, as presented in Figure 6. It can be observed that the customer, classified as having a "loss tendency" in the shipping behavior time series trend items exhibits a significant decrease in traffic volume over time.

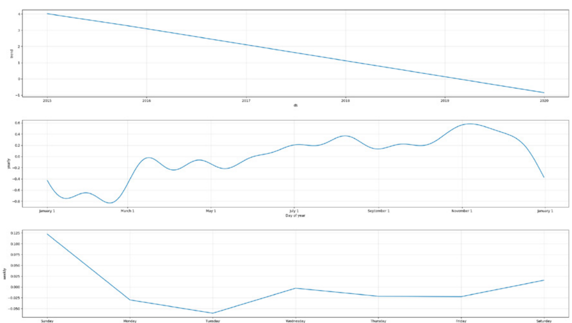

Figure 5. Traffic trend and cycle analysis for railway freight customer code 3196623

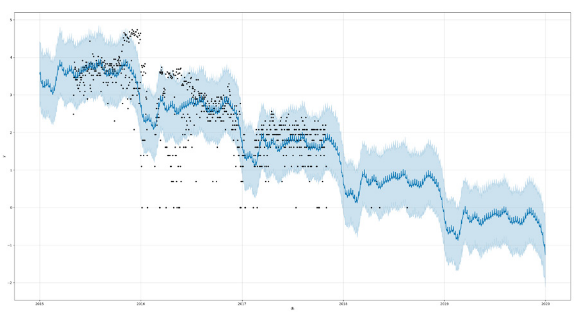

Figure 6. Traffic forecast for railway freight customer code 3196623

In addition to these behavioral characteristics, customers with a tendency to churn often exhibit distinct time series trends in their shipment volumes. These trends may include periodic fluctuations or volatility, but the underlying trend over time is characterized by a steady decline. By identifying and analyzing these trends, service providers can better understand the behavior of customers with a tendency to churn and develop effective strategies to mitigate this tendency, such as improving service quality or implementing targeted marketing campaigns.

5.3 Churn Classification Customer Characteristic Analysis

Lost railway freight customers are those who have ceased using rail freight services altogether, resulting in a complete loss of revenue for the service provider. These customers exhibit distinct behavioral characteristics, which can be identified and analyzed to better understand the factors contributing to their departure.

In this study, the loss characteristics of railway freight customers are analyzed, focusing on shipper code 3062229 as a case study. Figures 7 and 8 show the trend sequence and volume forecast sequence of this customer, respectively. The analysis revealed that the customer's shipping behavior was initially stable before July 2016, but showed a significant downward trend after that point, eventually leading to cessation. This pattern suggests that the customer was lost due to a change in his shipping needs, business priorities, or other factors.

Lost customers in the railway freight industry often exhibit unstable or erratic shipping behavior, with frequent fluctuations or significant drops in volume over time.

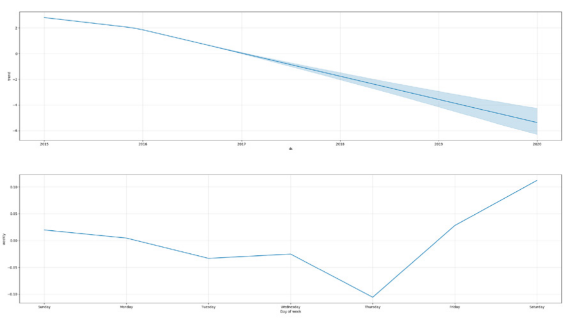

Figure 7. Traffic trend and cycle analysis for railway freight customer code 3062229

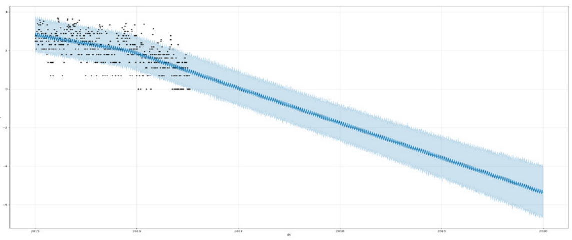

Figure 8. Traffic forecast for railway freight customer code 3062229

Identifying and analyzing the behavior of lost customers is critical for service providers looking to mitigate churn and maintain their customer base. By understanding the factors contributing to customer loss, service providers can develop targeted retention strategies and improve their overall service quality.

6. Conclusion

In the field of customer churn prediction, customer churn data often contain high-dimensional and time series features. Therefore, the developing of effective deep learning prediction models represents a crucial research topic. This paper proposes an improved LSTM algorithm to predict railway freight with time series characteristics and verifies its validity. The findings of the present research can be summarized as follows.

Firstly, the present approach utilizes the time series characteristics of railway freight customers to forecast, and the results of the present analysis demonstrate its efficacy.

Secondly, the loss tendency of customers is analyzed by using the time series prediction function of the improved LSTM algorithm, which provides technical support for railway enterprises to identify potential loss customers and improve their services.

Lastly, an intended future research will incorporate macroeconomic data and industry-related indicators that affect the freight industry. By comparing current transport demand with overall macro-industry trends, the uncovering of potential transport demand and loss tendencies is sought.

Overall, the proposed improved LSTM algorithm demonstrates promise in accurately predicting railway freight with time series characteristics, and the present analysis provides insights for the development of more effective churn prediction models in the future.

Acknowledgements

The research reported in this paper has been funded by the National Natural Science Foundation of China under Grant 61772065.

REFERENCES

Abduljabbar, R. L., Dia, H. & Tsai, P.-W. (2021) Unidirectional and Bidirectional LSTM Models for Short-Term Traffic Prediction. *Journal of Advanced Transportation*. 2021, 5589075. doi: 10.1155/2021/5589075.

Amorim, G. A., Lopes, L. A. S. & Silva Junior, O. S. (2020) Discrete Event-Based Railway Simulation Model for Eco-Efficiency Evaluation. *International Journal of Simulation Modelling*. 19(3), 375-386. doi: 10.2507/IJSIMM19-3-517.

Antoine, M., Ben Abdessalem, H. & Frasson, C. (2022) Cognitive Workload Assessment of Aircraft Pilots. *Journal of Behavioral and Brain Science*. 12, 474-484. doi: 10.4236/jbbs.2022.1210027.

Arikan, Y., Şen, T. & Çam, E. (2021) Energy Efficiency in Rail Systems with Coasting Control Method Using GA and ABC Optimizations. *Technical Gazette.* [*Tehnički Vjesnik*]. 28(4), 1127-1135. doi: 10.17559/TV-20200511115919.

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 (KDD ‘16), 13 – 17 August 2016, San Francisco, California, USA*. New York, USA, Association for Computing Machinery (ACM). pp.785-794. doi: 10.1145/2939672.2939785.

Chen, T., He, T. & Benesty, M. (2015) *Xgboost: Extreme Gradient Boosting*. *Package ‘xgboost’. R package version 0.4-2*. CRAN Repository. https://cran.microsoft.com/snapshot/2015-10-20/web/packages/xgboost/xgboost.pdf [Accessed 24th May 2023].

de Araujo, J. M. S. (2020) Combination of WRF Model and LSTM Network for Solar Radiation Forecasting – Timor Leste Case Study. *Computational Water, Energy, and Environmental Engineering*. 9, 108-144. doi: 10.4236/cweee.2020.94009.

De Caigny, A., Coussement, K. & Bock, K. W. D. (2018) A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*. 269(2), 760-772. doi: 10.1016/j.ejor.2018.02.009.

ICI Bucharest © Copyright 2012-2023. All rights reserved

Duan, G., Su, Y. & Fu, J. (2023) Landslide Displacement Prediction Based on Multivariate LSTM Model. *International Journal of Environmental Research and Public Health*. 20(2), 1167. doi: 10.3390/ijerph20021167.

Huang, J., Niu, G., Guan, H. & Song, S. (2023) Ultra-Short-Term Wind Power Prediction Based on LSTM with Loss Shrinkage Adam. *Energies*. 16(9), 3789. doi: 10.3390/en16093789.

Jailani, N. L M., Dhanasegaran, J. K., Alkawsi, G., Alkahtani, A. A., Phing, C. C., Baashar, Y., Capretz, L. F., Al-Shetwi, A. Q. & Tiong, S. K. (2023) Investigating the Power of LSTM-Based Models in Solar Energy Forecasting. *Processes*. 11(5), 1382. doi: 10.3390/pr11051382.

Jozefowicz, R., Zaremba, W. & Sutskever, I. (2015) An empirical exploration of recurrent network architectures. In: *Proceedings of the International Conference on Machine Learning, 6 – 11 July 2015, Lille, France*. New York, USA, Association for Computing Machinery (ACM). pp. 2342-2350.

Kasiran, Z., Ibrahim, Z. A., Syahir, M. & Ribuan, M. K. (2014) Customer Churn Prediction using Recurrent Neural Network with Reinforcement Learning Algorithm in Mobile Phone Users. *International Journal of Intelligent Information Processing*. 5(1), 1-11.

Krawiec, P., Junge, M. & Hesselbach, J. (2021) Comparison and Adaptation of Two Strategies for Anomaly Detection in Load Profiles Based on Methods from the Fields of Machine Learning and Statistics. *Open Journal of Energy Efficiency*. 10, 37-49. doi: 10.4236/ojee.2020.102003.

Li, X., Chen, H., Zhang, X. & Chen, W. (2018) A LSTM-based method for stock returns prediction: A case study of China stock market. In: Luo, F., Ogan, K., Zaki, M. J., Haas, L., Ooi, B. C., Kumar, V., Rachuri, S., Pyne, S., Ho, H., Hu, X., Yu, S., Hsiao, M. H-I. & Li, J. (eds.) *Proceedings of the 2015 IEEE International Conference on Big Data, 29 October – 1 November 2015, Santa Clara, United States*. New Jersey, USA, Institute of Electrical and Electronics Engineers (IEEE). pp. 2823-2824. doi: 10.1109/BigData.2015.7364089.

Mohana, S. D., Shiva Prakash, S. P. & Krinkin, K. (2023) Relationship LSTM Network for Prediction in Social Internet of Things. In: Kulkarni, A. J., Mirjalili, S. & Udgata, S. K. (eds.) *Intelligent Systems and Applications. Lecture Notes in Electrical Engineering*, 959, 133-141.

Nath, B., Kumbhar, C. & Khoa, B. T. (2022) A study on approaches to neural machine translation. *Journal of Logistics, Informatics and Service Science*. 9(3), 271-283. doi: 10.33168/LISS.2022.0319.

Obeidat, A. & Al-shalabi, M. (2022). An Efficient Approach towards Network Routing using Genetic Algorithm. *International Journal of Computers Communications & Control.* 17(5), 815. doi: 10.15837/IJCCC.2022.5.4815.

Obeidat, A. & Yaqbeh, R. (2022) Smart Approach for Botnet Detection Based on Network Traffic Analysis. *Journal of Electrical and Computer Engineering*. 2022, 3073932. doi: 10.1155/2022/3073932.

Soh, H., Sanner, S., White, M. & Jamieson, G. (2017). Deep sequential recommendation for personalized adaptive user interfaces. In: *Proceedings of the 22th International Conference on Intelligent User Interfaces, 13 – 16 March 2017, Limassol, Cyprus*. New York, USA, Association for Computing Machinery (ACM). pp. 589-593.

Stefenon, S. F., Seman, L. O., Aquino, L. S. & Coelho, L. D. S. (2023) Wavelet-Seq2Seq-LSTM with attention for time series forecasting of level of dams in hydroelectric power plants. *Energy*. 274, 127350. doi: 10.1016/j.energy.2023. 127350.

Xue, H., Liu, Y., Li, C. & Gao, F. (2022) Wheel Weighing Meter of Continuous Rail Based on BP Neural Network and Symmetric Moving Average Filter. *Technical Gazette. [Tehnički Vjesnik]*. 29(1), 278-284. doi: 10.17559/TV-20210824045232.

Yang, Y., Tian, N., Wang, Y. & Yuan, Z. (2022) A Parallel FP-Growth Mining Algorithm with Load Balancing Constraints for Traffic Crash Data. *International Journal of Computers Communications & Control.* 17(4), 4806. doi: 10.15837/ijccc.2022.4.4806.

Yang, Y., Yang, B., Yuan, Z., Meng, R. & Wang, Y. (2023a) Modeling and Comparing Two Modes of Sharing Parking Spots at Residential Area: Real-time and Fixed-time Allocation. *IET Intelligent Transport Systems.* 2023, 1-20. doi: 10.1049/itr2.12343.

Yang, Y., Yin, Y. X., Wang, Y. P., Meng, R. & Yuan, Z. (2023b) Modeling of Freeway Real-Time Traffic Crash Risk Based on Dynamic Traffic Flow Considering Temporal Effect Difference. *Journal of Transportation Engineering Part A: Systems*. 149(7), 04023063. doi: 10.1061/JTEPBS.TEENG-7717.

Yuan, Z., Yuan, X., Yang, Y., Jinjie Chen, J., Nie, Y., Cao, M. & Chen, L. (2023) Greenhouse Gas Emission Analysis and Measurement for Urban Rail Transit: A Review of Research Progress and Prospects. *Digital Transportation and Safety*. 2(1), 36-51. doi: 10.48130/DTS-2023-0004.

Zhang, B. (2019) *Research on the Theory and Method of Customer Churn Management in Railway Freight Transport.* PhD thesis. Southwest Jiaotong University.

Zhang, B., Peng, Q. & Liu, F. (2019) Research on customer churn prediction of railway scattered white goods based on parallel C4.5. *Computer Application Research*. 36(03), 829-832+837.

Zhang, Q. (2015) *Research on Customer Churn Analysis Model of Railway Freight Based on Business Intelligence*. PhD thesis. Southwest Jiaotong University.