ISSN 1533-9211

DOI 10.5281/zenodo.6891775

FRAUD AND LATE DELIVERY PREDICTION USING HYBRID MODEL

HEMAVATHI.R[1] and RAJAVARMAN.V.N[2]

[1]Research scholar, Department of Computer Science and Engineering,Dr. MGR Educational and Research Institute,Chennai,Tamilnadu, India. Email: hemavathy.cse@drmgrdu.ac.in

[2]Professor, Department of Computer Science and Engineering,Dr. MGR Educational and Research Institute,Chennai,Tamilnadu, India. Email: rajavarman.vn@drmgrdu.ac.in

ABSTRACT

The cargo sector is going through considerable expansion in volume owing to technical innovation in e-commerce and global trade liberalization. Volume expansion also indicates a surge in fraud cases involving smuggling and fraudulent reporting of goods. Shipping businesses and customs are largely dependent on normal random examination hence uncovering fraud is typically by coincidence. As the volume raises considerably it would no longer be viable and beneficial for both transportation firms and customs to pursue standard fraud detection tactics. Other related publications in this field have demonstrated that intelligent data-driven fraud detection is proved to be significantly more successful than regular inspections. The proposed system using machine learning algorithm for Support Vector Machine (SVM), Random Forest (RF) and Hybrid Scikit algorithms. As such in this article, we evaluate and then determine the most efficient methodologies and algorithms to detect fraud successfully within the shipping business. We also analyse characteristics that drive fraud activity, examine current fraud detection models, build the detection framework and apply the framework using the tool.

Keywords: Fraud detection Models, Support Vector Machine (SVM), Random Forest (RF) and Hybrid Scikit algorithms.

Introduction:

Nowadays, Enterprises have vast amounts of product data in the era of big data. Taking advantage of these data to evaluate the supply and demand situation effectively is a common issue faced by enterprises. Products fraud detection will enable companies to better understand the market and contribute to increasing the company's revenue. Therefore, this paper's research topic mainly focuses on the problem of DataCo supply chain fraud, which is provided by Kaggle competition.

Therefore, the objective of this article is to utilize the data of the company DataCo smart supply chain for the analysis of product fraud. And we propose a hybrid model, which output is the Arithmetic mean of the output of Rule Based Classification models. Compared with other models, this model has better advantages in performance. In the whole process, Python is the main tool for data processing, modeling, and analysis.

Objective:

In recent years, the rise of the Internet of things (IoT) as an emerging technology has been unbelievable, more companies are moving towards the adoption of these technologies

Many IoT sensors are being deployed to share information in real-time which eads to the generation of a huge amount of data. This data when used correctly, will be very helpful to the company to discover hidden patterns for better decision making in the future

This Research aims to **compare 3 popular machine learning classifiers** and measure their performance against neural network models to find out which machine learning model performs better.

The machine learning classifiers used in this project are

- Logistic Regression
- Support Vector Machines
- Random Forest classification

For fraud detection and to predict late delivery on the basis accuracy, recall score and F1 score.

System Analysis:

Existing System:

Regression is an important machine learning model for these kinds of problems. Predicting sales of a company needs time series data of that company and based on that data the model can predict the future sales of that company or product. So, in this research project we will analyze the time series sales data of a company and will predict the sales of the company for the coming quarter and for a specific product.

Proposed System:

Various authors have discussed predicting the significant features of sales prediction by using different machine learning and data mining techniques. We proposed a Logistic regression machine learning technique for sales prediction of significant features. After pre-processing the dataset Logistic Regression, a data mining classification technique was applied here by using the Sklearn library to analyze the score. Implementation of the Naïve Bayes method of getting accuracy results, and this classification results section done by using Python. Finally, at the end, compare the Comparing Model and Confusion Matrix results on the Logistic Regression algorithm. This classified data based on various organized features of sales dataset. Create a Logistic Regression model with the help of temporary variables and used the sigmoid function for graphical representation classified dataset.

Related Works:

Data mining refers to a process used by companies to turn raw massive data into useful information through algorithms and it depends on effective data collection, warehousing, and computer processing. In this task, the goal is to detect supply chain fraud, which extracts fraud patterns and knowledge from ordinary data. And then these fraud patterns can be used in further detection via clustering methods or classification methods.

To counter the problem more effectively, it is necessary to understand the technologies involved in detecting frauds.

Classification identifies:

The classification has been the most popular and the only way used so far to identify fraudulent financial statements [3]. Most financial statement fraud (FSF) auto-detection programs use supervised machine learning methodologies [4], which usually have a two-stage procedure, wherein the first stage a model is trained by using a training sample. In the second stage, objects are classified through the model obtained from the first stage. There are five methods include regression, decision trees, neural networks, Bayesian networks, and support vector machines.

Regression identifies:

Regression is the most widely used method to detect financial fraud [5]. Transformations of variables in regression models have also been studied in the context of fraud detection, including logit, stepwise-logistic, multi-criteria decision aid method, and exponential generalized beta two.

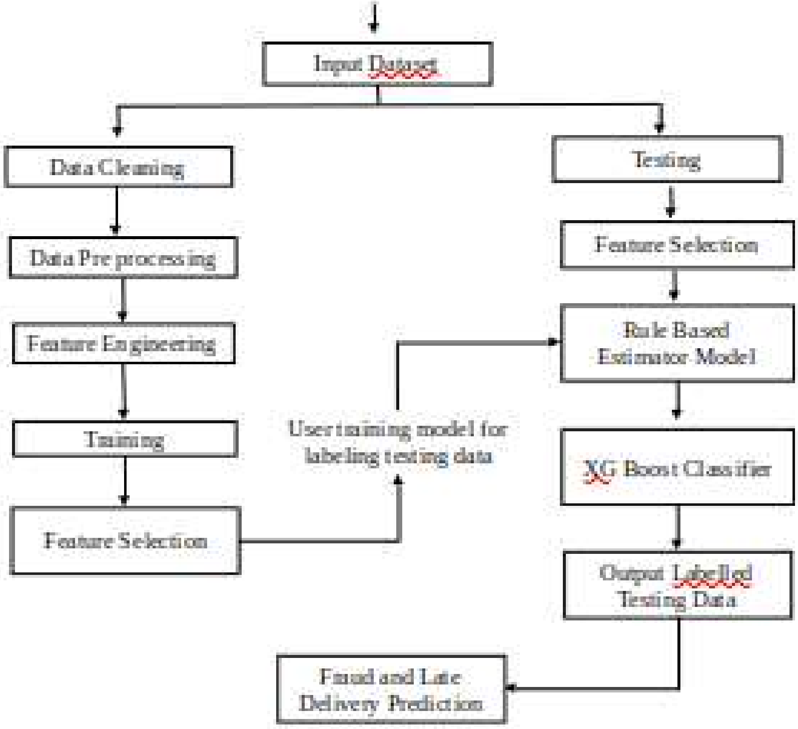
Neural network identify:

A neural network is another data mining technique that has been successfully used to detect financial statement fraud [6], neural network doesn't assume an attribute's independence and is capable of mining inter-correlated data and is a suitable alternative for problems where some of the assumptions associated with regression are not valid.

SYSTEM DESIGN:

Because of credit score card fraud activities there may be big economic losses. Trojan & phishing technologies used by the criminals to hack the records of other people’s credit card. Therefore, the fraud detection method is essential. Because of fraud detection approach we will identify a fraud in time when criminal makes use of fake card to patron. In this paper two kinds of random woodland set of rules are used to train the conduct characteristic of normal & fraud transactions. The aim of records analytics is to define hidden patterns &use them help informed conclusions in a variety of situations. Credit card fraud has exceptionally mismatch publicly to be had datasets. We pick out the most crucial variables which can cause better accuracy in credit card faux transaction detection.

DOI 10.5281/zenodo.6891775


Input Dataset
Data Cleaning
Data Pre processing
Feature Engineering
Training
Feature Selection
User training model for labeling testing data
Testing
Feature Selection
Rule Based Estimator Model
XG Boost Classifier
Output Labelled Testing Data
Fraud and Late Delivery Prediction

The range of transactions by credit score cards are increasing swiftly with the rapid improvement of digital commerce. The most popular transaction mode is on line buying, cases of transaction fraud is also growing. In these studies, we recommend a novel fraud detection technique that comprises of four tiers. First apply the cardholder's historical transaction facts to divide all cardholders into distinct businesses. Next, we summary a collection of particular behavioural patterns for each cardholder primarily based at the blended transactions & the cardholder's historical transactions. Then they educate a fixed of classifiers for each institution on the base of all behavioural patterns. Finally, to detect online fraud they use the classifier set.

In credit score card transactions detecting fraud is maybe one of the nice test beds for computational intelligence set of rules. This problem includes: Concept glide, elegance imbalance, verification latency. First, they advise, with the assist in their industrial companion, this device also reveal the most suitable performance measures for use for fraud detection functions. Second, we design an investigate a novel getting to know strategy that efficiently addresses elegance imbalance, concept glide & verification latency. Third, in this gadget exhibit the impact of sophistication unbalance & concept float in an actual-international records circulation.

When creating a credit card fraud detection version, it's far very essential to abstract the right features from transactional information. This is typically executed via combining the transactions so as to examine the outgoings behavioral patterns of the customers. In this

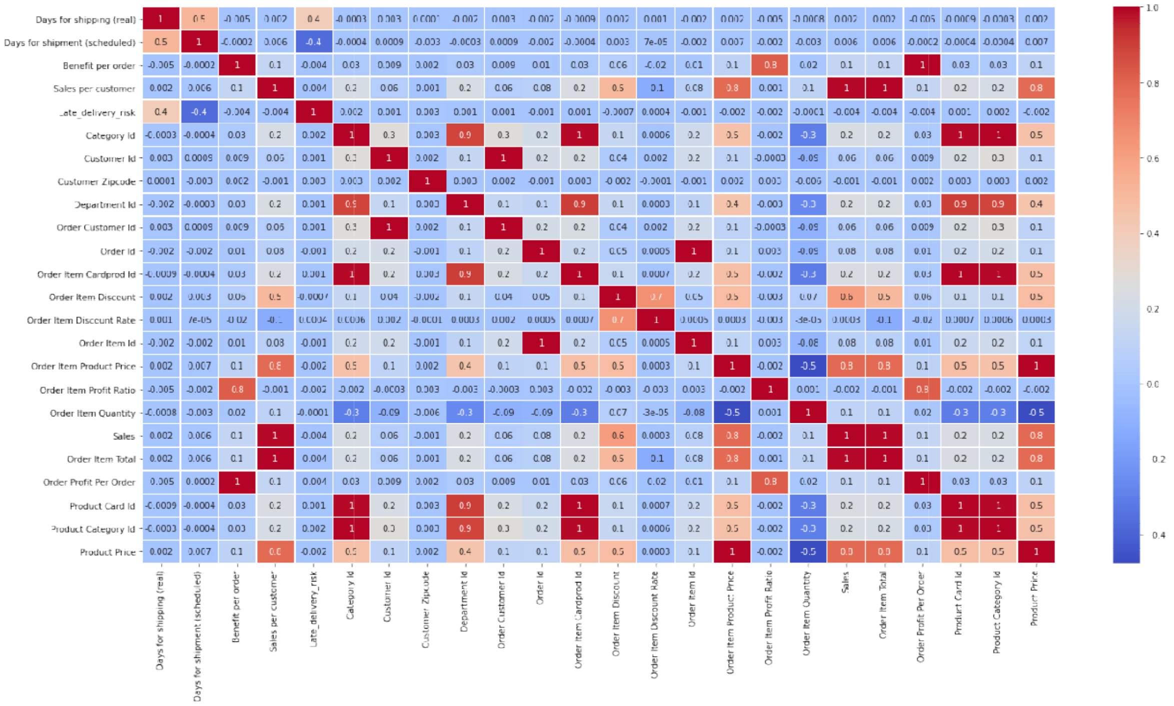
ISSN 1533-9211

DOI 10.5281/zenodo.6891775

gadget advice to create a brand-new set of functions based totally on analyzing the periodic behaviorof the time of a transaction the usage of the von Mises distribution. By which includes the proposed periodic functions into the techniques, the end result suggests a mean upward push in financial savings of 13%.

DATASET:

The dataset used in this project is maintained transparently with the Creative Commons 4.0 license by Fabian Constante, Fernando Silva, and António Pereira through the Mendeley data repository. For the prediction and analysis where it consists of 180519 rows and 53 columns in which the 16 columns are categorised as alphanumeric columns and remaining were the numerical columns. Below is the heatmap of the correlation matrix.



Dataset Description:

The types of products that need to be predicted are Clothing, Sports, Electronic Supplies and many more. Some of the attribute information are listed below.

Data Set Characteristics:	Multivariate	Number of Instances:	180519	Area:	Market Analysis
Attribute Characteristics:	Categorical, Integer, Real	Number of Attributes:	53	Date Donated	
Associated Tasks:	Classification	Missing Values?	Yes	Number of Web Hits:	

ATTRIBUTES TYPES:

- Real
- Ordered
- Binary
- Nominal

FEATURE ENGINEERING:

A Dataset of Supply Chains used by the company Data Co Global was used for the analysis. Areas of important registered activities are Provisioning, Production, Sales, and Commercial Distribution. It also allows the correlation of Structured Data with Unstructured Data for knowledge generation.

Independent Component Analysis:

Independent component analysis (ICA) is a widely-used blind source separation technique. ICA has been applied to many applications. ICA is usually utilized as a black box, without understanding its internal details. The basics of ICA are provided to show how it works to serve as a comprehensive source for researchers who are interested in this field by introducing the definition and underlying principles of ICA. Additionally, different numerical examples in a step-by-step approach are demonstrated to explain the pre-processing steps of ICA and the mixing and unfixing processes in ICA. Moreover, different ICA algorithms, challenges, and applications are presented.

The Seybold REPORT

ISSN 1533-9211

DOI 10.5281/zenodo.6891775

Principal Component Analysis:

Principal component analysis is one of the most important and powerful methods in chemo metrics as well as in a wealth of other areas. With a description of how to understand, use, and interpret principal component analysis- and also focuses on the use of principal component analysis in typical chemo metric areas but the results are generally applicable.

Exception Handling:

About the exception handling, we make sales of some goods that were shown as Nan in the tabular of the supply chain as 0.

Label Encoding:

We have converted the categorical columns into numeric using label encoding. The columns are Customer Country, Market, Type, Product Name, Customer Segment, Customer State, Order Region, Order City, Category Name, Customer City, Department Name, Order State, Shipping Mode, order_week_day, Order Country and Customer Full Name.

Design and Development

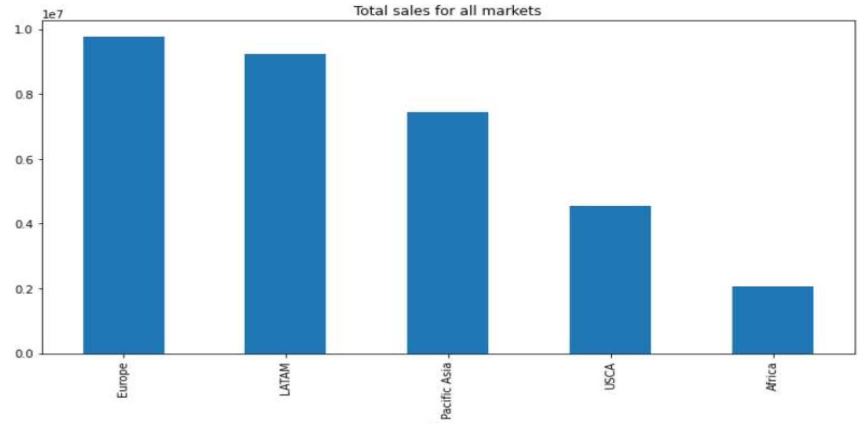
Sales Analysis:

Areas of important registered activities:

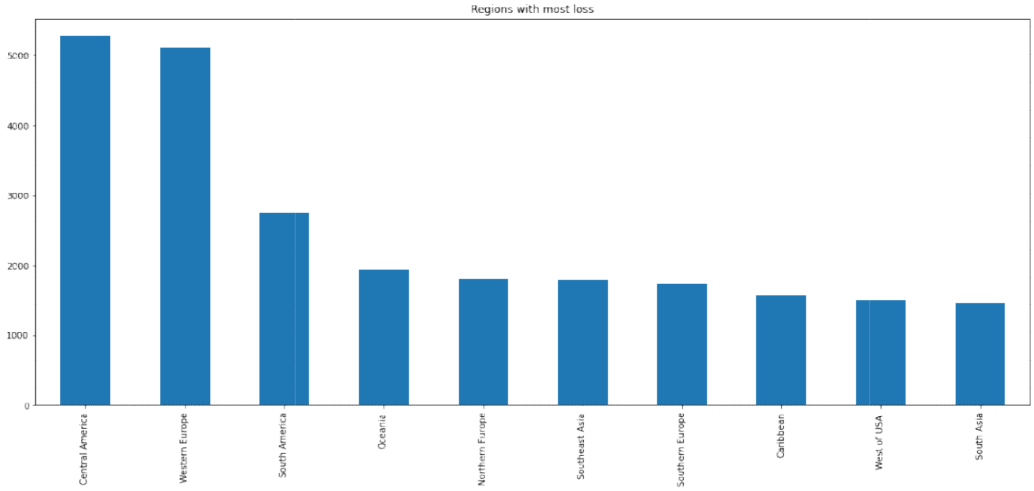
- Provisioning
- Production
- Sales
- Commercial Distribution

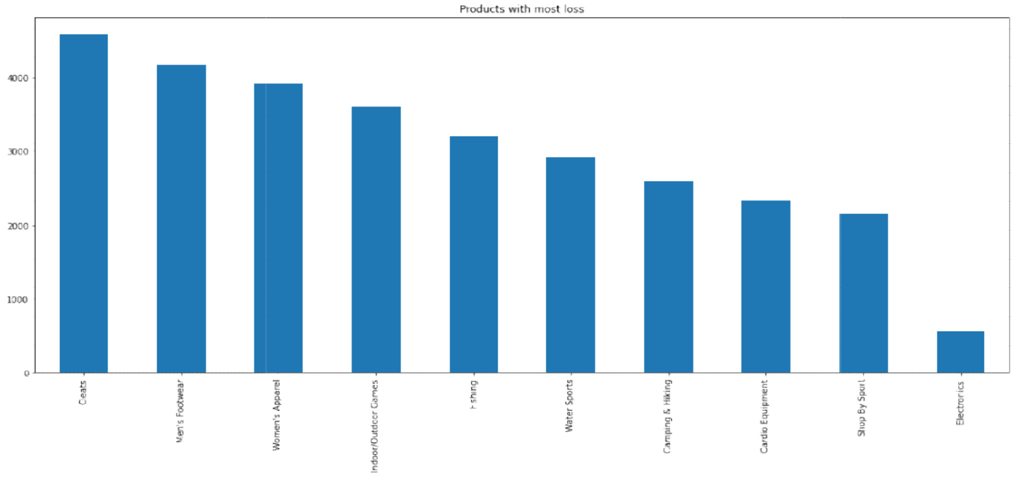
It also allows the correlation of Structured Data with Unstructured Data for knowledge generation.

From the Given data it is known that Europe has the highest of total sales from all the markets from which Western Europe has the most total sales referred from the given graph.



And also from the graphical representation of data it is found that Central America has the most loss and the products that faced the highest loss are Cleats and Men’s footwear.

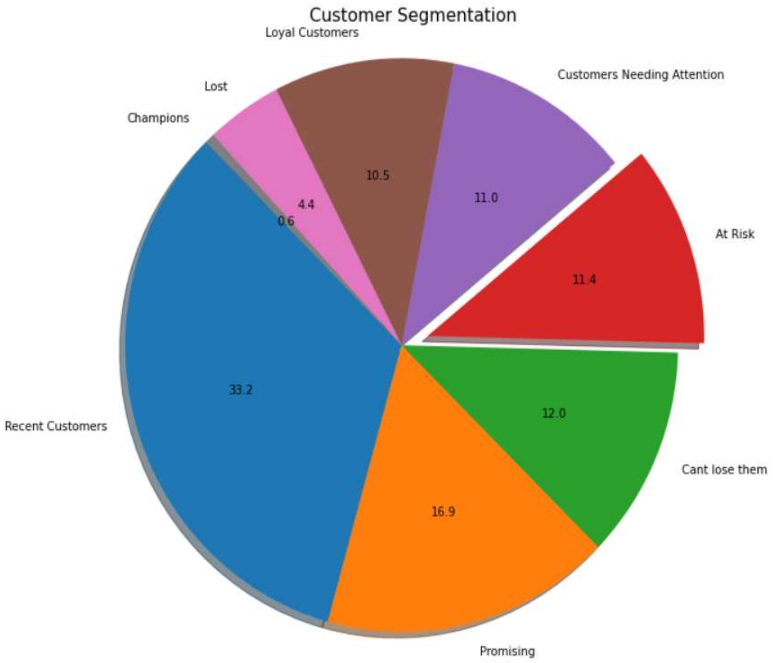




Customer Segmentation

DOI 10.5281/zenodo.6891775

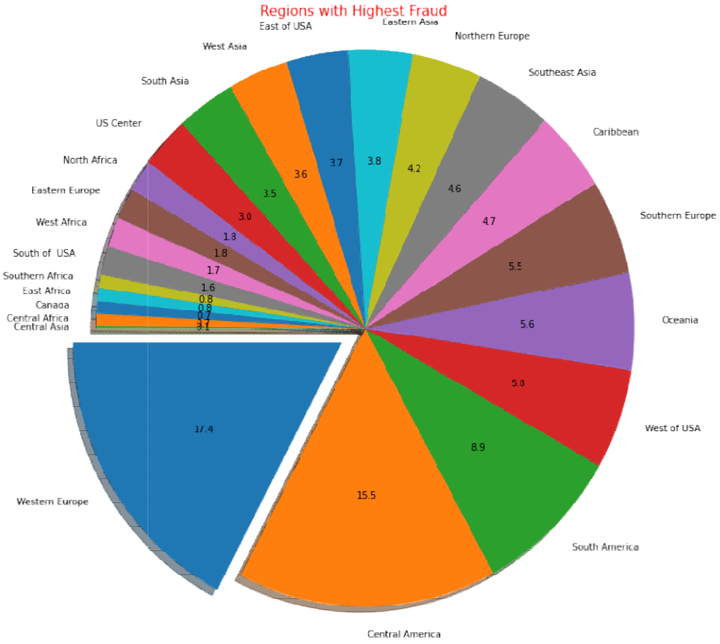
From the given data we simply infer that the count of promising customer is greater than the loyal customers whereas the count of recent customer is greater than the promising customers and the Champions are at the least which can be taken from the given pie chart.

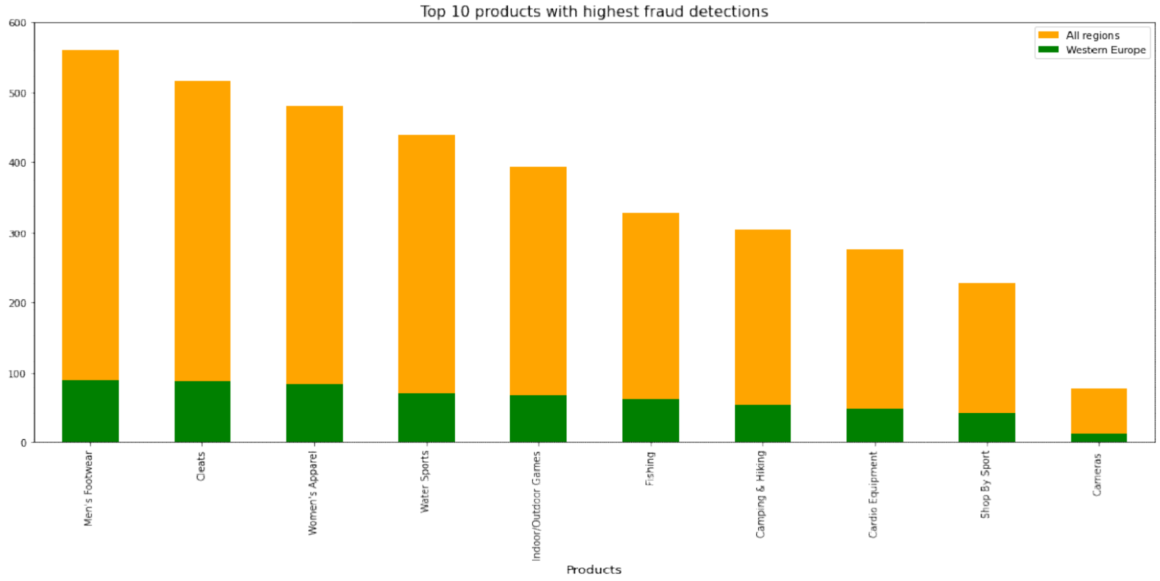


From the classification of input set dataset into seven kinds, which are Recent Customers, Champions, Lost, Loyal Customers, Customers Needing Attention, At-risk, Can't lose them. And the percent of them are 33.2%, 0.6%, 4.4%, 10.5%, 11.0%, 11.4%, 12.0% and 16.9% respectively.

Late Delivery and Fraud Detection:

Western Europe is resulted as the region with highest fraud count from the given pie chart representation of the most fraud count from the dataset used for predicton. And also Men's Footwear results to be the most frauded product where it shows as the highest fraud detection referred from the bar representation of top 10 products with highest fraud detections and cameras records to be the least.



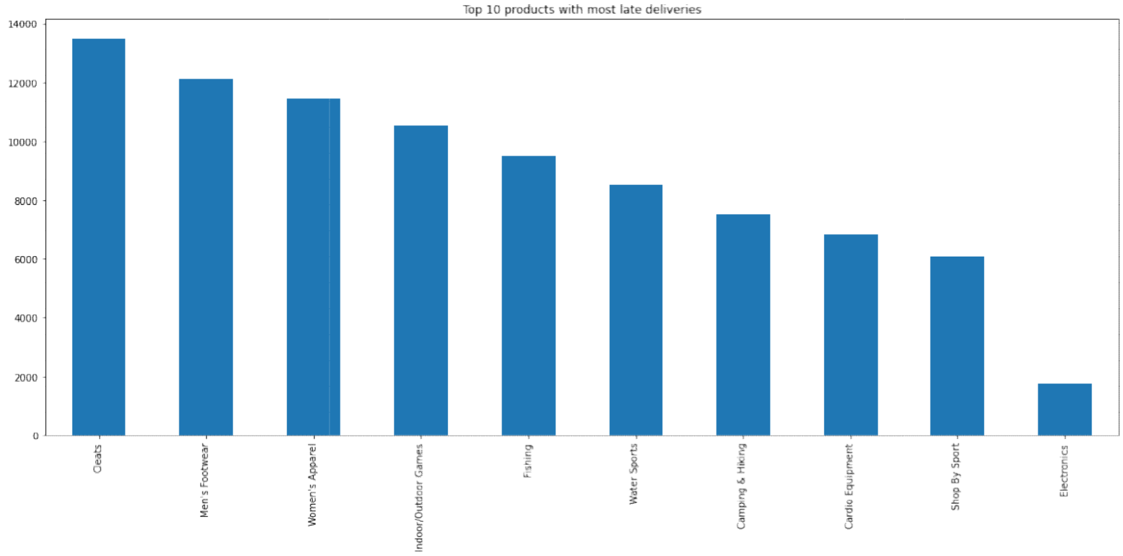


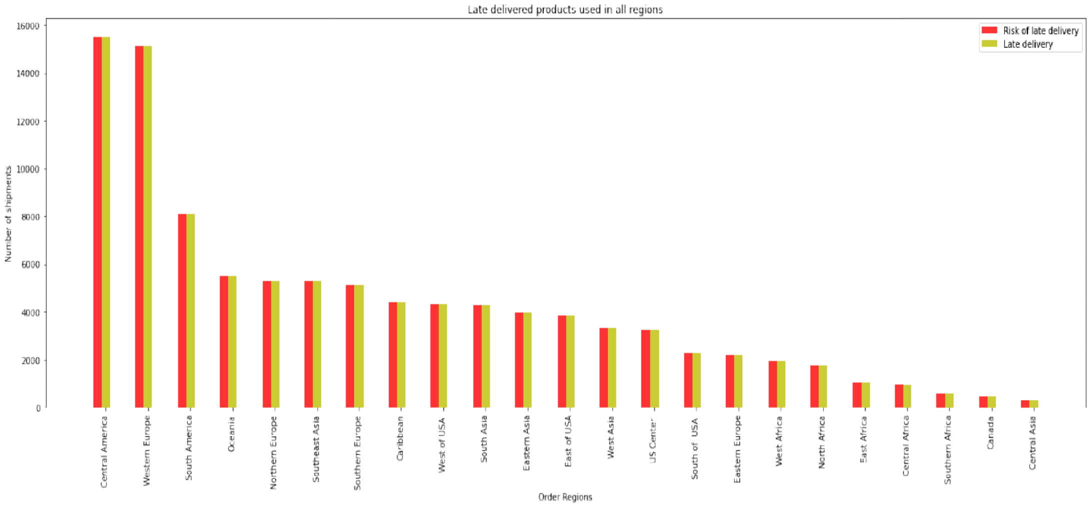
Electronics is the product with least late delivered product and Cleats recorded to be the highest with a count of more than 13500. It is also anticipated that Central America serves the top in both late delivery and risk in late delivery of products too and Central asia serves the least which is declared from the given bar representation from the given data.

The Seybold REPORT

ISSN 1533-9211

DOI 10.5281/zenodo.6891775





SUPPLY CHAIN FRAUD DETECTION MODEL

Machine Learning:

Machine learning is a new subfield in artificial intelligence. Its primary focus is on developing systems that can learn and forecast based on past experience.

The Importance of Domain Knowledge:

Any good machine learning project starts with domain knowledge collection, which is the process of obtaining relevant information and expertise about a business problem. Typically, we interact with industry practitioners, do online research, and execute data analysis to find unique trends, patterns, or indicators that may assist in the construction of machine learning models.

ISSN 1533-9211

DOI 10.5281/zenodo.6891775

Domain knowledge is particularly valuable for a multitude of reasons, including balancing stakeholder demands, knowing our target audience, and, most crucially, offering vital suggestions for feature engineering. While these indications are self-explanatory when attempting to identify a cat ina picture, feature engineering is far from intuitive in many businesses such as law, insurance, or medical diagnostics.

How to Create Rules Using Domain Knowledge

In many industries, simple deterministic rules may be derived from pre-existing processes. In some judicial procedures, for example, a claim for reputational damages may never be granted since the law code clearly says so. Because ML excels at resolving confusing and difficult instances, it makes sense to add deterministic rules into models only if they are relevant in all circumstances andare not too numerous or complicated. However, more use cases exist, therefore developed a comprehensive list of when you may want to consider installing hybrid, rule-based models:

- Deterministic rules for the probability model already exist
- Lack of data for particular sorts of prediction cases
- High feature number of occurrences
- Actively addressing biases in the data

Similarly, an insurance company may refuse to pay out damage claims less than $1,000 since they are not obligated to do so under the terms of their contract with the covered. If we intended to forecast, say, lawsuit results or insurance losses, such basic principles might be put straight into machine learning models to improve performance.

Classification Algorithms of machine learning:

✓ Random Forest

✓ Support Vector Machine

✓ Hybrid Rule-Based Machine Learning withScikit-Learn

A) Random Forest

The random forest approach may be utilized for classifications as well as regression applications. It enhances accuracy through cross validation. The random forest classifier will cope with incomplete data while keeping the accuracy of a considerable proportion of the data. It will not permit over-fitting trees in the model if there are extra trees. It is capable of processing big data sets with increasing diversity.

ISSN 1533-9211

DOI 10.5281/zenodo.6891775

Random Forest Algorithm:

Step 1: First, start with the identification of random samples from a particular dataset.

Step 2: Secondly, this technique will generate a decision tree for every sample. Then it will retrieve the forecast result from every tree structure.

Step 3: In this step, voting will be done for every expected outcome.

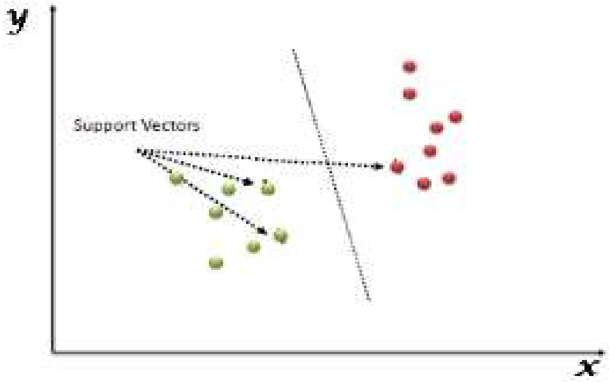
Step 4: At last, choose the highest voted forecast result as the's prediction result.

Pseudocode for Random Forest

```
Input:  dataset records
Training dataset T,
T= (t1, t2, t3,tn) in testing dataset. Predictor variable value
Output: record
A class of testing dataset.
1) Read the training dataset as T
2) Calculate the Precision and f1-score of the predictor variables in each class
3) Repeat Calculate the probability of T, using the Multimodal RF in each class then Until the probability of all predictor variables (t1, t2, t3,. .... tn)has been calculated.
4) Calculate the probability for each class and classify the text from the attributes
5) Get the maximum probability.
6) Predict the output using y_pred and y_test utility
```

B) Support Vector Machine (SVM)

SVM is a supervised learning approach that may be used to regression and classification problems. Each data item is represented by a single in n-dimensional space in the SVM technique



Pseudocode for

Support Vector Machine Algorithm:

Input: dataset records

1. Import the required packages.

2. Convert the string values in the dataset to numerical values.

3. Assign the data to X_train, y_train, X_test and y_test variables.

4. Using train_test_split() function, pass the training and testing variables and give test_size and the random_state as parameters.

5. Import the SVClassifier from sklearn library.

6. Using SVClassifier, predict the output of the testing data.

7. Calculate the accuracy

C) Hybrid Rule-Based Model

Use Domain Knowledge to Enhance Scikit-learn Models with Hard-Coded Rules Supervised machine learning models are excellent at making predictions in the face of uncertainty; they detect patterns in historical data and reliably project them into the future. ML has pushed the boundaries in sectors where finding the most probable result, whether a class or a particular number, has traditionally proven difficult, error-prone, or too time-consuming or costly at scale. In such a setting, it seems inefficient to have an ML model use implicit learning to asymptotically estimate pre-formulated rules. Instead, we want the model to concentrate on all scenarios in which no pre-defined rules exist. In this post, you will learn about the several advantages of embedding pre-defined domain rules into machine learning models. To get further hands-on experience, we will create a simple wrapper class for scikit-learn estimators that takes explicit rules into account while leaving the model to handle the difficult scenarios.

How to Hard-Code Deterministic Rules as Logical Formulae

As previously stated, ML models learn rules implicitly. Such learning is shown by decision-tree-based algorithms such as scikit-Decision learn's Tree Classifier or Gradient Boosting Regressor, which is an ensemble of decision trees. Decision-tree-based algorithms seek to forecast the target variable by learning decision rules based on the data provided. The decision rules themselves are really simple; they are a series of splits on the data that use just the fundamental logical operators =>. Nonetheless, all splits merely approximate any clear restrictions and, as a result, may be less precise. Using the same method, we may create basicdeterministic rules as logical formulas, which we can then convert into code. For example, imaginewe wish to create a predictive model to forecast an insurance company's overall losses, and we know the firm rejects claims worth less than $1,000. This rule might be hard-coded in one of two ways:

```
Ifclaim_amount<= 1000:
# reject claim else:
# use machine learning model
```

A Standard Rule Format

We format rules as a Python dictionary, as seen above. The feature column names to which we wish to apply our rules are represented by the dictionary keys. The dictionary's values are lists of tuples, with each tuple representing a distinct rule.

Pseudocode: Hybrid Rule-Based XGBOOST processing

```
Input: D: Micro-blog dataset
N: The number of trees in XGBOOST
Output: Discrete data
Load actual dataset
Divide into subsets
Set $d_s$ = gap start date
Set $d_e$ = gap end date
$M( x_0 )$
$x_1$ , $M ( x_0 ), x_2$
fori = 0; i< D; i++ do
preprocessing the data text participle
end for
while LDA does not converge do
extracting text subject characteristics
end while
Run the model and get predictions
$y_1 = M_1$ (gap)
$y_2 = M_2$ (gap)
fori = 0; i< C; i++ do
end for
```

The logical operator of the rule is the first member of the tuple, the split criteria is the second, and the last object is the value that the model should provide if the rule is applied. For example, the first rule in the above example would state that the model should return 0.0 if any valuein the House Price feature column is less than 1000.0.

Other Necessary Methods

We need to create two more basic methods get params and set params to get a functional model that inherits from the Base Estimator class. We may use them to set and read the parameters of our new model.XGBoost is a newly dominant technique in applied ML and Kagglecontests for structured or tabular data. XGBoost is a gradient-boosted decision tree implementation optimised for speed and performance. In this post, you will learn about XGBoost and receive a light introduction to what it is, where it comes from, and how you can study more.XGBoost is a decision-tree-based ensemble Machine Learning technique that employs a gradient boosting framework. In prediction issues involving unstructured data (pictures, text, etc.), artificial neural networks surpass all existing algorithms or platforms. However, when it comes to specific structured data, decision tree-based algorithms are now regarded best-in-class. Please read the graphic below to observe how tree-based algorithms have evolved throughout time.

Hybrid Rule-Based Gradient Boost Model Usage Example

Here is a short code snippet to illustrate how you could utilize the Rule Augmented Estimator wrapper class to add rules to a Gradient Boosting Classifier. This example assumes you have already initialized the variables rules, train_X, train_y, and test_X. Please refer to the section A Common Format for Rules to examine how any rules should be employed. XGboost has proven to be the most efficient Scalable Tree Boosting Method. The system runs way faster on a single machine than any other machine learning technique with efficient data and memory handling.

CONCLUSION:

In this paper, we propose a hybrid model to detect the supply chain fraud of the DataCoCompany. We use the Rule Based Classification on DataCO smart supply chain dataset provided by the Kaggle competition platform. Experiments show the Confusion Martixes for Random Forest and Rule Based Classification, and the quantity of TP and TF are huge, which means our model performs well. Furthermore, our method can effectively my features of different dimensions and performs better than the other algorithms and the SVM.

References:

[1] N. Mahmoudi, E. Duman, "Detecting credit card fraud by Modified Fisher Discriminant Analysis", Elsevier Expert System with Application, 2015, pp. 2510-2516.

[2] M. Zareapoor, K. Seeja, M. Alam, "Analysis of credit cardfraud detection techniques: based on certain design criteria", International Journal Compututer Application, 2012, pp. 35–42.

[3] V. Vlasselaer, C. Bravo, O. Caelen, T. Eliassi-Rad, L. Akoglu, M. Snoeck, B. Baesens, "APATE: A Novel Approach for Automated Credit Card Transaction Fraud Detection using Network based Extensions", ELSEVIER Decision Support Systems, 2015, pp. 38-48.

ISSN 1533-9211

DOI 10.5281/zenodo.6891775

[4] A. Bahnsen, D. Aouada, A. Stojanovic, B. Ottersten, “Feature Engineering Strategies for Credit Card Fraud Detection”, ELSEVIER Expert System with Applications, 2016, pp. 134-142.

[5] J. Quah, M. Shriganesh, “Real-time credit card fraud detection using Computational Intelligence”, Expert System Application, 2008, pp. 1721-1732.

[7] Yue, X. Wu, Y. Wang, Y. Li, C. Chu, “A review of data mining-based financial fraud detection research”, international conference on wireless communications Sep, Networking and Mobile Computing, 2007, pp. 5519–5522.

[8] Chen, Tianqi&Guestrin, Carlos. (2016). “XGBoost: A Scalable Tree Boosting System”, 785-794. 10.1145/2939672.2939785. [2] Chen, T., He, T., Benesty, M., et al.: Xgboost: extreme gradient boosting. R package version 04-2, 1-4 (2015)

[9] Y. Dianmin, W. Xiaodan, W. Yunfeng, L. Yue, C. Chao-HsienA review of data mining-based financial fraud detection research International Conference on Wireless Communications, Networking and Mobile Computing (2007), pp. 5519-5522

[10] M.S. BeasleyAn empirical analysis of the relation between the board ofdirector composition and financial statement fraud, The Accounting

Review of Finance, 71 (4) (1996), pp. 443-465

[11] T.B. Bell, J.V. Carcello, A Decision aid for assessing the likelihood offraudulent financial reporting, Auditing, 19 (1) (2000), pp. 169-184

[12] W.-S. Chen, Y.-K. Du, Using neural networks and data mining techniquesfor the financial distress prediction model, Expert Systems with Applications Part 2, 36 (2) (March 2009), pp. 4075-4086

[14] Goodenough, John B. "Exception handling: issues and a proposed notation." Communications of the ACM 18.12 (1975): 683-696.

[15] Ellis, Katherine, et al. "A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers." Physiological measurement 35.11 (2014): 2191.

[16] Valecha, Harsh, et al. "Prediction of consumer behaviour using random forest algorithm." 2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). IEEE, 2018.

[17] Visa,Sofia,etal."Confusion Selection." MAICS 710 (2011): 120-127. Matrix-based Feature

[18] Marom, Nadav David, LiorRokach, and Armin Shmilovici. "Using the confusion matrix for improving ensemble classifiers." 2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel. IEEE, 2010.

[19] Chicco, Davide, and Giuseppe Jurman. "The advantages of the Matthewscorrelation coefficient (MCC) over F1 score and accuracy in binaryclassification evaluation." BMC genomics 21.1 (2020): 6.

[20] Meyer, David, Friedrich Leisch, and Kurt Hornik. "The support vectormachine under test." Neurocomputing 55.1-2 (2003): 169-186.

[21] Bôhning, Dankmar. "Multinomial logistic regression algorithm." Annalsof the institute of Statistical Mathematics 44.1 (1992): 197-200.

[22] Griffis, Joseph C., Jane B. Allendorfer, and Jerzy P. Szaflarski. "Voxel-based Gaussian naïve Bayes classification of ischemic stroke lesions inindividual T1-weighted MRI scans." Journal of neurosciencemethods 257 (2016): 97-108.