Fraud Detections
&
Warranty Claims

Team: Raghav Jindal,Tejaswi Dabas

AGENDA

- Introduction
- Goal
- Materials and methods used
- Procedure
- Results

Introduction

Data mining is the process of discovering patterns in large datasets to make informed decisions. This project uses data mining techniques to detect fraud in warranty claims. The goal is to develop a machine-learning model that accurately classifies claims as genuine or fraudulent based on historical data. The project applies various techniques, including Generalized Linear Models and Support Vector Machines, to identify important features and build an accurate predictive model. The project also discusses limitations and provides recommendations for future research. The outcome aims to help the company identify and reject fraudulent claims, improving its financial stability and maintaining the trust of its genuine customers.

Data Mining Goal

This project aims to develop a machine-learning model that can accurately classify warranty claims into genuine and fraudulent categories. The model should be able to analyze various features associated with each claim, such as product information, customer information, and claim details, to identify patterns that indicate potential fraud.

About the Dataset

This page is having the data related to an organization with 20 variables. In that 19 are independent variables to predict the dependent variable 'fraud' with 2 categories '1' indicates fraudulent claim and '0' indicates genuine claim. This data is having 8000 about observations.

Data-Fields

- Region - Customer region details
- state - Current location of customer
- Area - Area_Urban/rural
- City- Customers current located city
- Consumer_profile- Customer's work profile
- Product_category- Product category
- Product_type- Type of the product_Tv/Ac
- AC_1001_Issue- 1001 is failure of Compressor in AC
- AC_1002_Issue- 1002 is failure of Condenser Coil in AC
- AC_1003_Issue- 1003 is failure of Evaporator Coil in AC
- TV_2001_Issue- 2001 is failure of power supply in Tv
- TV_2002_Issue- 2002 is failure of Inverter in Tv
- TV_2003_Issue- 2003 is failure of Motherboard in Tv
- claim_value- Customer's claim amount in Rs
- Service_Centre- 7 Different service centers
- Product_Age- Duration of the product purchased by customer
- Purchased_from- From where product is purchased
- Call_details- call duration in mins
- Purpose- Purpose_compliant-Compliant raised by customer claim- claimed for the product Other- Other categories out of this
- Fraud- '1'- fraudulent claim, '0' Genuine claim

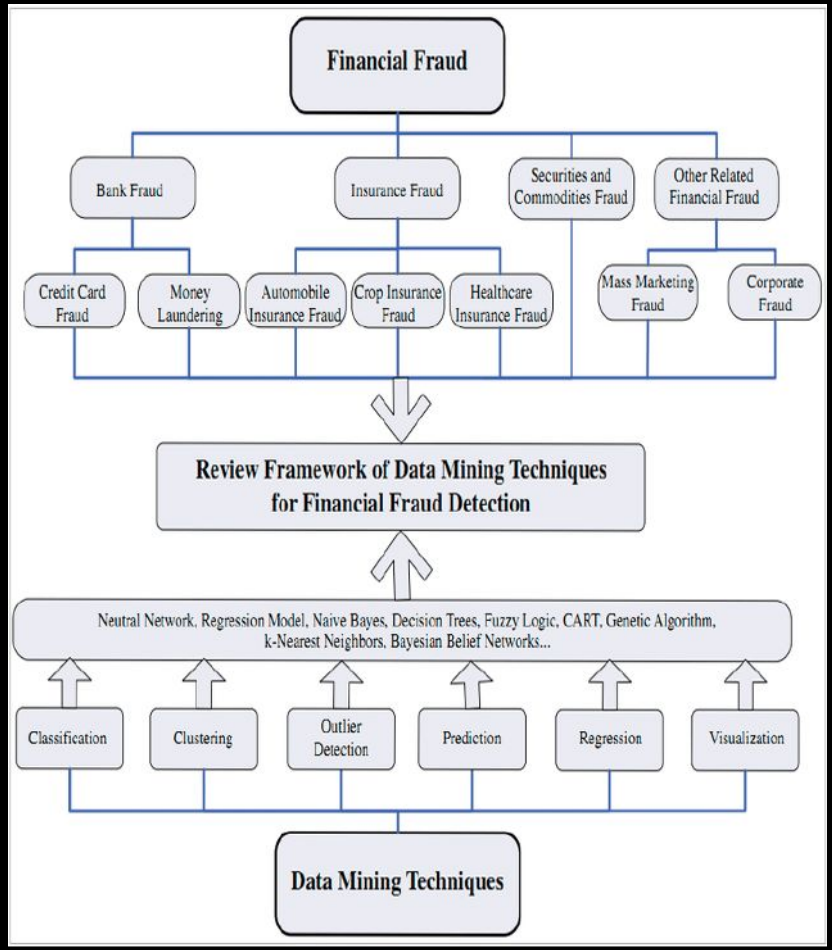
Method Used

About Data Set:

The dataset for this project consists of historical warranty claims, including information about the product, customer, and type of claim filed. It also includes specific issues reported for air conditioner and TV products, as well as details about the location of the claim, amount claimed, age of the product, and more. The target variable is a binary column indicating whether the warranty claim was fraudulent or genuine.

Data Cleaning:

To ensure accurate data analysis, it's important to have a clean data collection. Data cleansing is a crucial stage in this process, involving reviewing and examining data for mistakes, inconsistencies, and inaccuracies. Techniques for cleansing data include handling missing values through imputing or removal, and standardizing data to ensure uniformity.



Algorithm Used

1. GLM (Generalized Linear Models) is a statistical model used for regression analysis that allows the response variable to have non-normal error distributions.
2. SVM (Support Vector Machines) is a machine learning algorithm used for classification and regression analysis that creates a hyperplane to separate classes.
3. Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions to improve accuracy.
4. KNN (K-Nearest Neighbors) is a classification algorithm that predicts the class of a new observation based on the class of its k nearest neighbors.
5. CART (Classification and Regression Trees) is a decision tree algorithm that recursively splits the data based on the best feature to create a set of rules to predict the target variable.
6. GBM (Gradient Boosting Machines) is an ensemble learning method that builds multiple decision trees sequentially, where each subsequent tree tries to correct the errors of the previous tree.
7. Neural Network is a machine learning algorithm inspired by the structure of the human brain that uses layers of interconnected nodes to learn and make predictions.
8. Weka is a collection of machine learning algorithms for data mining tasks, including classification, regression, clustering, and feature selection, implemented in Java and available through a graphical user interface. R has an interface to Weka through the RWeka package.

Attributes selection method used

1. Random Forest: An ensemble learning method that constructs a multitude of decision trees and combines them to improve accuracy and prevent overfitting.
2. Correlation-based feature selection: A method that evaluates the correlation between each feature and the target variable, and selects the most highly correlated features.
3. Recursive feature elimination (RFE): A method that recursively removes the least important features and selects the remaining ones based on a specified algorithm.
4. RFECV: Recursive feature elimination with cross-validation, which uses cross-validation to determine the optimal number of features to select.
5. RFE-SVM: A combination of RFE and Support Vector Machines (SVM) for feature selection.
6. PCA: Principal Component Analysis, a method for reducing the dimensionality of high-dimensional datasets by identifying the most important features that explain the variance in the data.
7. Chi-squared attribute selection: A method that calculates the chi-squared statistic between each feature and the target variable and selects the features with the highest statistic.

Attributes selected by each selection method

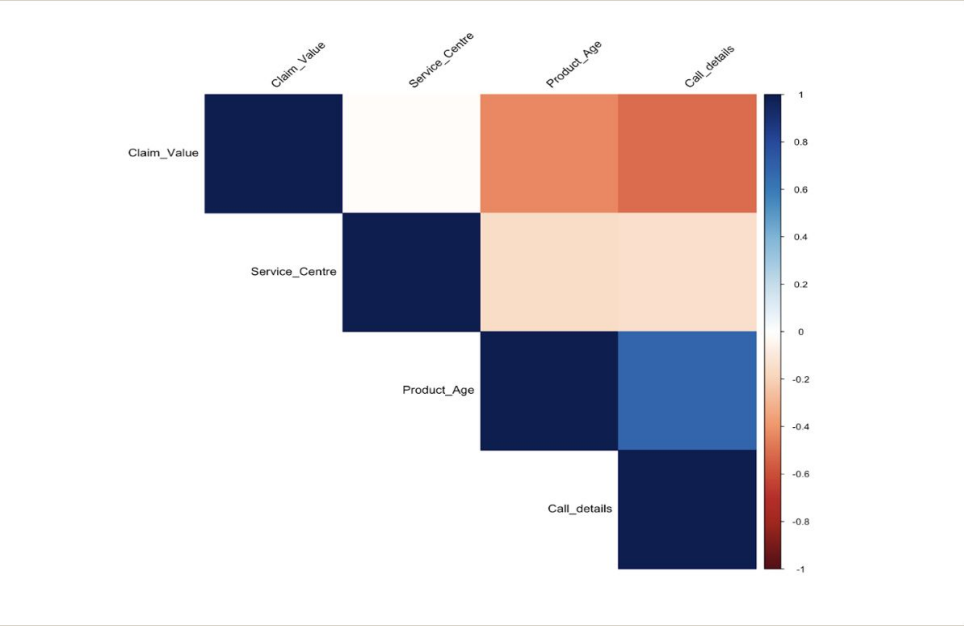
1. **Recursive Feature Elimination (RFE):**
   Fraud,Purchased_from,City,Product_Age,Call_details,Claim_Value,State,Service_Centre,Purpose
2. **Recursive Feature Elimination with Cross-Validation (RFECV):**
   Fraud,Purchased_from,City,Product_Age,Call_details,Claim_Value,State,Service_Centre,Purpose,TV_2002_Issue,Consumer_profile,Area,TV_2003_Issue,AC_1001_Issue,AC_1002_Issue,Product_type,Product_category,TV_2001_Issue,AC_1003_Issue
3. **Recursive Feature Elimination with Support Vector Machine (RFE-SVM):**
   Fraud,Purchased_from,City,Product_Age,Call_details,Claim_Value,State,Service_Centre,Purpose,TV_2002_Issue,Consumer_profile,Area,TV_2003_Issue,AC_1001_Issue,AC_1002_Issue,Product_type,Product_category,TV_2001_Issue,AC_1003_Issue
4. **Principal Component Analysis (PCA):**
   State,Area,City,Consumer_profile,Product_category,Product_type,Purchased_from,Purpose,Fraud
5. **Random Forest:**
   Fraud,Product_Age,Claim_Value,City,Call_details,Purchased_from,State,Service_Centre,Purpose
6. **Filter-based method using the chi-square:**
   Fraud,Product_Age,Claim_Value,Call_details,Purchased_from,City,State,Service_Centre,AC_1002_Issue
7. **Correlation-based:**
   Claim_Value,TV_2002_Issue,TV_2003_Issue,Service_Centre,Fraud

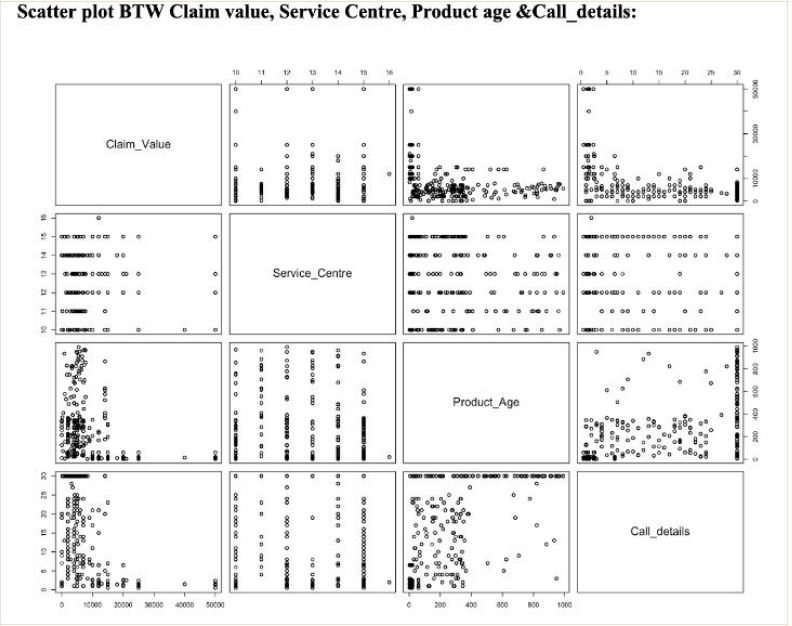
Procedure

1. Defining the categorization methods we would use, such as Naive-Bayes, Weka, Neural Net, Support Machine Vector, and Random Forest, was the first stage in the Analysis portion of the R code. The next stage entailed applying 10-fold cross-validation using these eight classification algorithms to the full dataset. Then, in order to assess the efficiency of the categorization algorithms, we gathered the average measures for accuracy, true positive rate, false positive rate, precision, recall, F1-measure, Matthew's correlation coefficient (MCC), and ROC AUC for the 10 folds. After data pre-processing, the clean dataset only had rows and the original dataset had about 8000 rows and 20 characteristics. This suggests that the original sample contained a sizable number of missing values and/or anomalies.
2. By eliminating anomalies and null values during the pre-processing stages, the machine learning models that were taught on the data performed better. This indicates that these actions were successful in raising the dataset's quality.
3. To determine the probability that a person will go back to jail after being released, eight machine learning models were chosen: Weka, Random Forest, Support Vector Machine (SVM), Neural Nets, and Naive Bayes. This suggests that the issue being addressed was a categorization issue.
4. Using a 66-33 divide to separate the dataset into training and test data, the models were trained on the training data and assessed on the test data using a confusion matrix. This indicates that a holdout technique was used to correctly validate the models.
5. Of all the models,Random Forest did the best at predicting the probability that a person would go back to prison after being released, with an accuracy rate of 97.6% across the entire dataset.
6. To extract sections of the most crucial characteristics from the original dataset, several attribute selection methods were used:Correlation Based Feature Selection, Random Forest Feature Selection, and Recursive Feature Selection and few more.The most pertinent characteristics to include in the models were determined using these attribute selection methods, which can enhance the performance of the models.

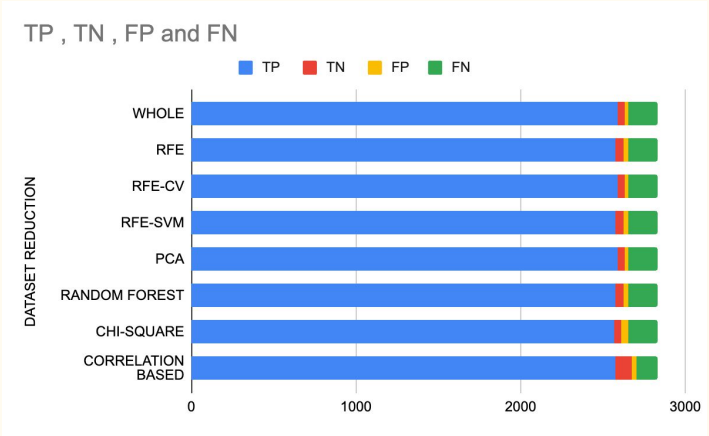
The data were divided into training and testing groups for each of the segments derived from the attribute selection methods, and the eight machine learning models were trained on the newly created training dataset. 64 models in total were taught and assessed.

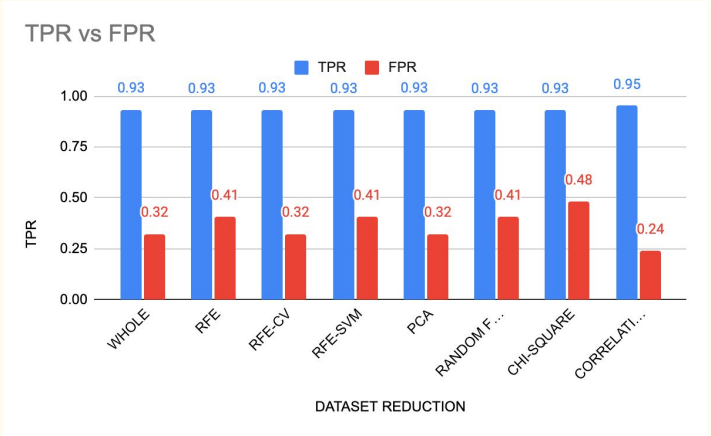
Data Visualization

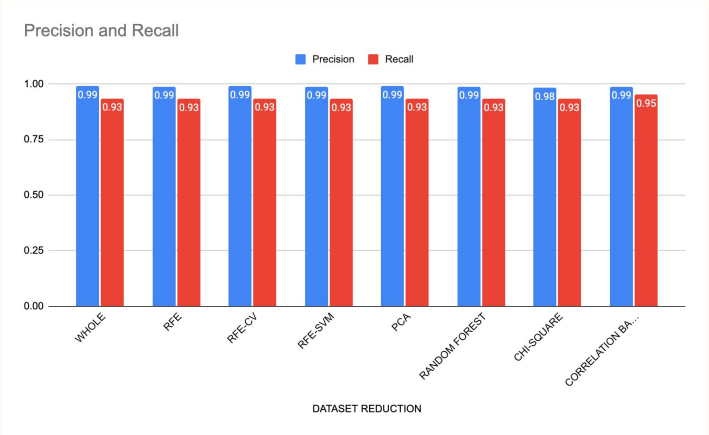


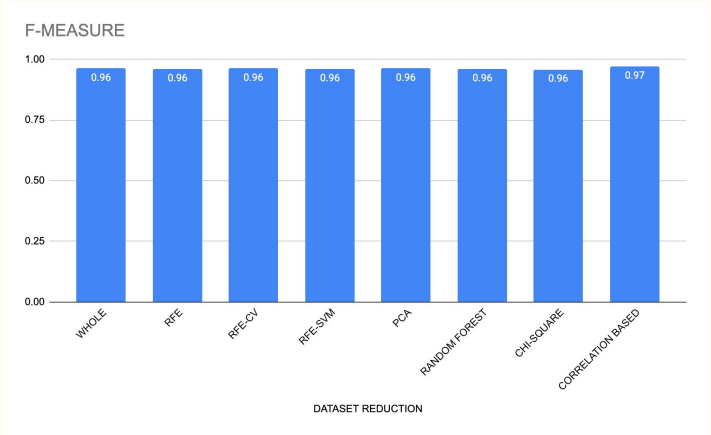


Graphs

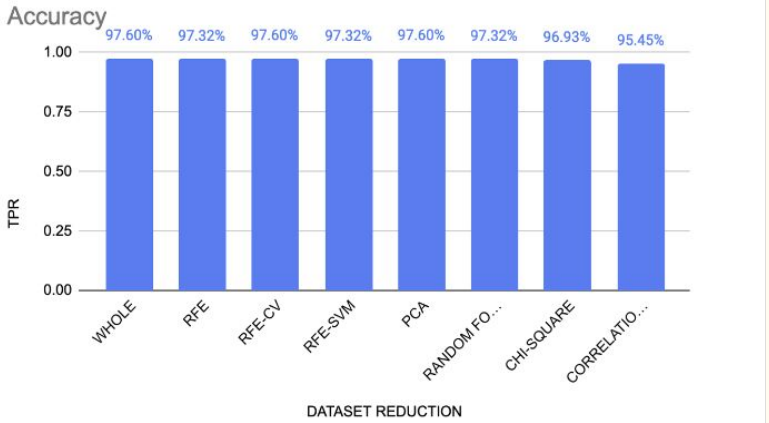

TP , TN , FP and FN
TP
TN
FP
FN
WHOLE
RFE
RFE-CV
RFE-SVM
PCA
RANDOM FOREST
CHI-SQUARE
CORRELATION BASED
DATASET REDUCTION
0
1000
2000
3000


TPR vs FPR
TPR
FPR
1.00
0.75
0.50
0.25
0.00
TPR
0.93
0.93
0.93
0.93
0.93
0.93
0.93
0.95
0.32
0.41
0.32
0.41
0.32
0.41
0.48
0.24
WHOLE
RFE
RFE-CV
RFE-SVM
PCA
RANDOM F…
CHI-SQUARE
CORRELATI…
DATASET REDUCTION


Precision and Recall
Precision
Recall
1.00
0.75
0.50
0.25
0.00
0.99
0.93
0.98
0.95
WHOLE
RFE
RFE-CV
RFE-SVM
PCA
RANDOM FOREST
CHI-SQUARE
CORRELATION BA…
DATASET REDUCTION


F-MEASURE
1.00
0.75
0.50
0.25
0.00
0.96
0.97
WHOLE
RFE
RFE-CV
RFE-SVM
PCA
RANDOM FOREST
CHI-SQUARE
CORRELATION BASED
DATASET REDUCTION

Result



BEST PERFORMANCE MEASURES WITH ACCORDANCE TO EACH REDUCTION ATTRIBUTE SELECTION

DATASET REDUCTION	MODEL	TP	TN	FP	FN	Accuracy (R)	TPR	FPR	Precision	Recall	F-Measure
WHOLE	RF	2587	46	22	180	0.976	0.9349475967	0.3235294118	0.9915676504	0.9349475967	0.9624255952
RFE	RF1	2578	45	31	181	0.9732	0.9343965205	0.4078947368	0.9881180529	0.9343965205	0.9605067064
RFE-CV	RF2	2587	46	22	180	0.976	0.9349475967	0.3235294118	0.9915676504	0.9349475967	0.9624255952
RFE-SVM	RF3	2578	45	31	181	0.9732	0.9343965205	0.4078947368	0.9881180529	0.9343965205	0.9605067064
PCA	RF4	2587	46	22	180	0.976	0.9349475967	0.3235294118	0.9915676504	0.9349475967	0.9624255952
RANDOM FOREST	RF5	2578	45	31	181	0.9732	0.9343965205	0.4078947368	0.9881180529	0.9343965205	0.9605067064
CHI-SQUARE	RF6	2567	45	42	181	0.9693	0.9341339156	0.4827586207	0.9839018781	0.9341339156	0.9583722233
CORRELATION BASED	RF7	2578	98	31	128	0.9545	0.9526977088	0.2403100775	0.9881180529	0.9526977088	0.970084666

Best Model: Random Forest Classifier with 97.6% for reduced dataset according to Recursive Feature Elimination with Cross Validation, Principal Component Analysis feature reduction

Conclusion

The conclusion drawn from the project highlights several important aspects of machine learning. Firstly, the significance of feature selection in machine learning is emphasized. Feature selection techniques enable the identification of relevant features that contribute most to the prediction task, which can reduce the time taken to train machine learning models, while improving their performance.

The project also demonstrates how proficiency in training machine learning models can be improved using the R programming language.The project's experience in using R for machine learning can be useful for future projects, as it enables efficient data manipulation, feature selection, and model development, which can significantly enhance the performance of machine learning models.

Furthermore, the project has provided a deeper understanding of the strengths and limitations of different classification models and how to optimize their performance. By comparing and contrasting the performance of eight different machine learning algorithms, including KNN, Filter-based method using the chi-squar**e,** Random Forest, Support Vector Machine (SVM), Neural Nets, and the team gained insights into the trade-offs between different algorithms and how to choose the most suitable algorithm for a specific task.

In conclusion, the project has provided valuable practical experience in various aspects of machine learning, including data pre-processing, feature selection, model selection, and evaluation, using the R programming language. The team gained a deeper understanding of the strengths and limitations of different classification models and how to optimize their performance, which can be useful for future machine-learning projects in diverse domains.