

ENSEMBLE MODELING & PREDICTION INTERPRETABILITY FOR INSURANCE
FRAUD CLAIMS CLASSIFICATION

Dissertation submitted in part fulfilment of the requirements for the degree of

Master's in Data Analytics

At Dublin Business School

MADHANA VEERAPANDIAN BALASUBRAMANIAN

Student ID: 10365017



Master's in Data Analytics

January, 2019

DECLARATION:

I, Madhana Veerapandian Balasubramanian, declare that this research is my original work and that it has never been presented to any institution or university for the award of Degree or Diploma. In addition, I have referenced correctly all the literature and sources used in this work and this work is fully compliant with the Dublin Business School's academic honesty policy.

Signed: Madhana Veerapandian Balasubramanian

Date: 07/01/2019

ACKNOWLEDGEMENT:

This dissertation would not have been possible without the constant support and encouragement of many people.

First and foremost, I would like to thank God for giving me the power, understanding, talent and chance to take on this project and complete it satisfactorily. Without his blessings, this achievement would not have been possible.

I would like to thank my father, my mother and my brother for always supporting, encouraging and trusting me in everything I have chosen to do in my life. Also, I would like to thank my friends for their encouragement and support.

In my journey towards this degree, I offer my sincerest gratitude to my supervisor Dr. Shahram Azizi Sazi for sharing his constant motivation, advice and valuable guidance throughout the term. I consider this opportunity as one of the most important experience of my life.

In my journey towards my career, I would like to thank my mentor Mr. Martin Brown for his support and motivation for me in pursuing my interest in Fraud/Risk analytics using Machine Learning.

Also, Big thanks to the people of Ireland for the hospitality and meetup communities around Dublin where I gained lot of industrial exposure and formed a network.

ABSTRACT

The insurance fraud claims classification using Ensemble modeling is explained in this research paper. Using the pattern found in the data, Machine learning algorithms were able to find the fraud claims efficiently. The goal of this research is to carry out ensemble models like Gradient Boosting Machine, Random Forest and XGBOOST algorithms with sampling techniques and compare the results obtained with the traditional algorithms like SVM, Logistic Regression and Artificial Neural Networks. This research used data produced by Oracle and classifiers were trained on features selected after feature engineering with Boruta package in R, Chi-Square test, Sample T-test and evaluated with metrics such as Accuracy, ROC and F1 score. The result showed that ensemble methods meet high ROC score than traditional methods. XGBoost algorithm achieved highest AUC score of about 86.8% after over sampling using SMOTE on training data. Local Interpretable Model-Agnostic Explanations (LIME) package was used for model interpretability which gave a good insight on individual prediction. Also, a prediction API was developed using plumber package in R.

CONTENT

1	INTRODUCTION	1
	1.3.1.3 UNDER SAMPLING	4
	1.4 FEATURE ENGINEERING.....	4
	1.4.1 CHI-SQUARE TEST.....	5
	1.4.2 SAMPLE T-TEST.....	5
	1.4.3 BORUTA PACKAGE IN R.....	6
	1.6 EVALUATION AND INTERPRETABILITY	12
	1.6.1. METRICS EVALUATION USING CROSS-VALIDATION (10-FOLDS).....	12
	1.6.2 MONTE-CARLO SIMULATION.....	12
	1.6.3 LEARNING CURVE	13
	1.7 MODEL INTERPRETABILITY	13
	1.7.1 LIME.....	14
	1.7.2 SHAPLEY	14
	1.7.3 DALEX	15
	1.8 RESTFUL API FOR PREDICTION.....	15
	1.8.1 API USING PLUMBER IN R.....	16
	1.8.2 API USING FLASK IN PYTHON.....	17
	1.9 RESEARCH OBJECTIVE.....	17
	1.10 RESEARCH QUESTION.....	18
2	LITERATURE REVIEW	19
3	RESEARCH METHODOLOGY	23
	3.1. DATA DESCRIPTION	23
	3.2 METHODOLOGY.....	24
	3.2.1 H2O FRAMEWORK.....	24
	3.2.2 DATA VISUALIZATION	24
	3.2.3 FEATURE ENGINEERING	25
	3.2.3.1 INDEPENDENT SAMPLE T-TEST AND CHI-SQUARE TEST:.....	25
	3.2.4 DATA PRE-PROCESSING AND DATA PREPARATION:	26
	3.3.2 RAPIDMINER AUTO ML FRAMEWORK.....	28
	3.4 TRADITIONAL & ENSEMBLE MACHINE LEARNING ALGORITHMS	28
	3.5 EVALUATION METRICS	29

3.5.1 ACCURACY PARADOX.....	29
3.5.2 ACCURACY	29
3.5.3 ROC CURVE.....	30
3.5.3 PRECISION AND RECALL.....	31
3.5.4 F1 SCORE	32
3.6 LIME.....	32
3.7 PLUMBER API.....	32
4 DATA ANALYSIS/FINDINGS AND DISCUSSION	33
4.1 BORUTA FEATURE SECTION.....	33
4.2 AUTO ML	35
4.3 ENSEMBLE MODELS	37
4.4 OPTIMAL TREE DEPTH ANALYSIS FOR XGBoost:	37
4.5 RESULTS EXPLANATION	38
4.6 INTERPRETABILITY OF BLACK BOX MODELS	40
4.7 REST API USING PLUMBER.....	40
5 CONCLUSIONS AND FUTURE WORK	41
5.1 REFLECTIONS.....	41
6 BIBLIOGRAPHY	43

LIST OF ABBREVIATIONS

XGBOOST	Extreme Gradient Boosting
ROC	Receiver Operating Characteristics
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
GBM	Gradient Boosting Machine
DT	Decision Tree
LIME	Local Interpretability Model Agnostic Explanations
ROS	Random Over Sampling
AUTO ML	Automated Machine Learning

LIST OF FIGURES

<i>Figure 1 Feature explanation</i>	23
<i>Figure 2 Distribution of label</i>	25
<i>Figure 3 Receiver Operating Characteristics</i>	31
<i>Figure 4 Boruta variable explanation boxplot</i>	33
<i>Figure 5 Receiver Operating Characteristics from RapidMiner Auto Model</i>	36
<i>Figure 6 Metrics plot from RapidMiner Auto Model</i>	36
<i>Figure 7 Learning curves for models</i>	37
<i>Figure 8 Optimal tree depth for XGBoost</i>	38
<i>Figure 9 LIME prediction for XGBoost</i>	40
<i>Figure 10 API output from postman app</i>	40

LIST OF TABLES

<i>Table 1 HTTP request methods [27]</i>	16
<i>Table 2 P-value from Chi-Square</i>	26
<i>Table 3 Confusion Matrix</i>	29
<i>Table 4 Boruta Variable Importance</i>	34
<i>Table 5 Auto ML Leader Board</i>	35
<i>Table 6 Metric table for built models</i>	39

1 INTRODUCTION

Insurance fraud claim is one of the many fraud activities that challenges many domains like Banking, Health and Motor insurance. It is estimated that Insurance fraud activities will increase in the coming days and costs many to the industries based on the researches published until now. There are various researches conducted in this space to reduce the fraud activity using advanced data analytics. Machine Learning and its sub-field Deep learning are being used in this field to find the fraudulent activity and stop it. Increase in data and reduction in computing costs allows us in identifying trends and patterns efficiently to help companies to improve their relationship with clients, process optimization, resource administration and increase the profits. Deployment is the biggest challenge after developing predictive models for mitigating fraud. Deployment is one of the most crucial tasks for the organizations. Fraud costs the insurance industries in America an estimated value of \$80 billion a year [1] and in developing countries the estimated cost \$600 million a year, of which claims fraud is a substantial contributor. Many of the Big Data Science implementations on insurance claims has been investigated but focused on medical insurance and not short-term insurance like motor insurance [2][3]. As of the published papers, Motor insurance needs more attention and research in fighting frauds at this stage to avoid financial loss and to reduce the steady increase in cost of insurance to all customers.

1.1. FRAUD MODELING

Insurance fraud is “the wrongful or criminal deception of an insurance company for the purpose of wrongfully receiving compensation or benefits.” [4] Insurance fraud can be further divided into planned and opportunistic fraud or hard and soft fraud [5]. False accidents/injuries created by criminals are part of planned fraud called as hard fraud. The policy holders of insurance self-inflate their claims to increase the amount is seen as opportunistic/soft fraud. Our research is focused on the claims made by the fraudsters. Though, there is lot of researches and implementation on fraud modeling and mitigation of fraud claims, Fraudsters are coming up with a new approach and there is a serious need of machine learning techniques to reduce the illegal claims to learn the unforeseen patterns/anomaly in the claim data.

Therefore, frauds in these systems are considered as cyber-crime since it causes a huge amount of financial losses to the companies and increases the overall insurance cover.

There are two main problems in data mining-based fraud detection research

1. Lack of publicly available real data in this domain to do experiments. There were various challenges in obtaining proper access to financial observations to do research in this area and is very hard owing to **privacy and competitive** reasons.

2. Lack of published well-researched methods and techniques for fighting and resolving the issue.

1.2. AUTO INSURANCE FRAUD CLAIMS

Auto Insurance fraud claims are one of the areas where lot of illegal claims are made with false accidents/injuries. A big problem exists in insurance companies related to this situation that are dealing with a fraud giant presented in auto insurance in a research. Organizations are losing millions of dollars and it is driving them to find suitable methodologies with machine learning models to have solution against fraudulent activities [20].

Few actions made by the fraudsters in the name of claim are given as follow

- Claiming the same bill to various insurance companies.
- Service Bills which cannot be justified for the insurance cover.
- Vehicle insurance policy is used in charging medical treatments.
- Health insurance companies which makes fake bills on nonexistent patients and applies unnecessary treatments like surgeries to the injured person who really is not.
- Fake bills charge to nonexistent health entities in the insurance companies.
- Inflation of treatments and medicines costs owing to the use of fake policies to claim.
- Forming a gang and setting up a story to make the claim real.
- Link between authority and fraudster in claiming fraud bills

1.3. IMBALANCED DATASETS

Dataset for fraudulent claims are always imbalanced and it is a serious problem for modeling and evaluation of the machine learning model.

Imbalanced data may greatly affect the performance of classification algorithms. The prediction will be biased towards the majority class present in the dataset. Therefore, sampling methods should be employed to solve the data imbalance problem. Because there is a large difference in the amount of data between the classes of claims.

1.3.1 SAMPLING STRATEGIES

We can both under sample legitimate claims (majority class) and oversample fraudulent claims (minority class) to balance the dataset for preparing the dataset. We can also use techniques like SMOTE (Synthetic Minority Oversampling technique), SOMO (Self Organizing Maps based Oversampling technique) to over/under sample fraudulent claims and randomly under sample legitimate claims to get the same amount of data from the majority class to form a balanced data set. "To minimize imbalance-biased estimates of performance, one of the researches advised the use of reporting both the resulting metric values and the degree of imbalance in the data" [6].

The goal of this kind of approach applied before any classification algorithm is to reduce the skewed distribution that exist in the data either by introducing synthetic individuals to the minority class (resampling) or deleting instances from the majority class (under sampling). There were various different methods that have been proposed to reach a balanced data based by adding new individuals belonging to the minority class [7].

1.3.1.1 RANDOM OVERSAMPLING

Random oversampling technique is the easiest to equalize the data distribution between classes by taking individuals from the minority class and replicating them based on the skewness present. This improves

the performance of the algorithms and ROS is an independent algorithm for sampling which mimics the datapoints exactly as it is present in the real.

1.3.1.2. SMOTE

SMOTE was found to deal with the previous enumerated by randomly selecting each time an instance from the minority class and identify some of the nearest neighbors of it, based on the Euclidean distance, and create new individuals based on 6 linear interpolations between the selected item and its neighbors, it is important to remark that the overlap between the two cases can be increased due to the generation of the same number of artificial instances for each minority individual.[8]

1.3.1.3 UNDER SAMPLING

When we sample the non-fraud instances randomly to match the fraud instances which is 923 in our dataset, we form a under sampled dataset to match the count of non-fraud instances. In this formation we can measure the accuracy of the model. The disadvantage of under sampling is that we must lose the number of instances in overall dataset which will suppress the data quality of removing valid instances that can be used for the modeling.

1.4 FEATURE ENGINEERING

Machine Learning (ML) requires features (i.e., attributes/variables) to train the model to predict/classify an objective. One of the key challenges for modeling is to determine the correct number and the type of such features from the overall features available in the dataset. We can use all available features in the dataset for modeling, but it will be prone to overfitting, predictive errors, bias and poor generalization. So, we need to measure the weight of each feature in predicting/classifying out of all features available and check if it has either a unique predictive value, redundant, or irrelevant value. The key to better accuracy and building model using ML is to identify the set of right features [9].

There are various techniques available for feature engineering corresponding to the type of features we are trying to reduce. Predominantly, Principal Component Analysis, Linear Discriminant Analysis and other clustering algorithms are employed to reduce huge number attributes to some meaningful attributes. As

far as the dataset is concerned, Statistical feature engineering techniques such as Chi-Square and T-test can be done to understand the variable importance.

1.4.1 CHI-SQUARE TEST

The chi-square test is a statistical independence test for checking if two categorical variables are related and how it is significant to each other.

The Chi-square statistical test is a test to analyze group differences when the dependent variable is measured at a nominal level which is non-parametric in nature. The result from the test allows the researcher to get the insights and derive more detailed information from the statistics than other tests/techniques. The use of this test is its robustness with respect to distribution of the data, easiness in computation, the details that can be fetched from the results are helpful where parametric assumptions cannot be met and flexibility in handling data from both two group and multiple group studies. There are some limitations such as sample size needs, difficulty in interpreting when categories (20 or more) in the independent or dependent variables are large and tendency of the Cramer's V to produce relative low correlation measures even for highly significant results. [10]

1.4.2 SAMPLE T-TEST

Independent samples t-test is useful in determining whether there is a statistically significant difference between the means in two unrelated groups. We can have a threshold p-value to accept or reject the variable. This test consists of One independent categorical variable that has two levels/groups and one continuous dependent variable.

In Statistics, there are 2 hypotheses for any analysis

1. Null hypothesis can be given as $H_0: \mu_1 = \mu_2$
2. Alternate hypothesis can be given as $H_A: \mu_1 \neq \mu_2$

Based on the significance value, we can reject/accept the feature.

1.4.3 BORUTA PACKAGE IN R

Boruta is an all relevant feature selection wrapper algorithm suitable for any classification method that gives important features as output which we can use for modeling. It uses Random Forest algorithms and does top-down search for relevant features by comparing original feature importance with significance at random which is calculated using permutation technique and elimination of features occur simultaneously based on the importance to stabilize the test [11]. It takes exact copy of the features and removes one by one by building the feature importance for all the features finally coming up with the best features based on the probability score of each feature. This package is very much influential in feature selection in recent researches for compressing the features.

1.5 TRADITIONAL AND ENSEMBLE MODELS

1.5.1 TRADITIONAL ML MODELS

In Supervised Machine Learning, there are 2 major division as namely Regression and Classification algorithms. The ML algorithms which are used commonly are named as Traditional models which are Support Vector Machines, Logistic Regression, Decision Trees and Artificial Neural Networks.

1.5.1.1 SUPPORT VECTOR MACHINES

Support Vector Machine is a discriminative classifier. A classification algorithm that tends to maximize the margin between positive and negative classes by projecting input data vectors to a higher dimensional space. This algorithm depends on the support vectors sitting on the plane and are affected only by points near the margin for any classification task it handles. SVM works very well with high-dimensional data which is bit challenging comparatively for other algorithms.

SVM CLASSIFICATION:

$$\min_{f, \xi_i} \|f\|_K^2 + C \sum_{i=1}^l \xi_i \quad \mathbf{f}(\mathbf{x}_i) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

SVM CLASSIFICATION, DUAL FORMULATION:

$$\min_{\alpha_i} \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \quad \mathbf{0} \leq \alpha_i \leq C, \text{ for all } i \quad \sum_{i=1}^l \alpha_i y_i = 0$$

ξ_i calculated the error made at particular point (\mathbf{x}_i, y_i) . [12]

There are other variations which solves non-linear data problem unlike simple SVM which is good at linearly separated classes and it is called as Kernel SVM.

The different kernels available are

1. Polynomial Kernel
2. Gaussian Kernel('RBF')
3. Sigmoid Kernel

We can use any of this based on the dataset we have.

1.5.1.2 LOGISTIC REGRESSION

Logistic regression is another discriminative classifier which is used as interpretable algorithm in solving many Machine learning based problems. It is one of the most widely used learning algorithms in classification for its interpretability.

The logistic function used in this algorithm which is also called as **sigmoid function**, a S-shaped curve that can accept any real-valued number and assign it into a value between 0 and 1, but never exactly at those limits.

The coefficients must be estimated from the data points using maximum-likelihood estimation which is a common learning algorithm used by a many ML algorithm though it always makes assumptions about the distribution of your data.

The logistic function is given as below

$$\text{logit}(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_kX_k$$

The logistic regression coefficients are $b_0, b_1, b_2, \dots, b_k$

$$p = \frac{1}{1 + e^{-\text{logit}(p)}}$$

p-probability [13]

Other logistic regression forms are

1. Binary Logistic Regression
2. Multinomial Logistic Regression
3. Ordinal Logistic Regression

1.5.1.3 ARTIFICIAL NEURAL NETWORKS

A neural network (NN) consists of many simple connected processors called neurons which involves in producing a sequence of real-valued activations for the network. There are two ways of activating neurons, one through sensors in the environment and other through weighted connections from previously active neurons. From these activations, some neurons may influence the environment by triggering actions for next phase. Based on the problem and connections between neurons, the influence may require very long causal chains of computational steps, where each step transforms (often in a non-linear way) the aggregate activation of the network for further process. [41]

All categorical, numeric and topical features are transmitted into the input layer of the deep neural network (DNN) to start the training process. Then, the input layer maps the features to the first hidden layer, and the process continues. Each hidden layer includes several nodes for processing the input data

of the layer and transporting the result to the next layer. The activation function of each layer can add non-linear mapping to the mapping process to guarantee that the abstraction ability of the DNN is more effective. At the same time, to avoid the gradient vanish of errors in the process of back propagation, this paper employed ReLU instead of Sigmoid as the activation function. After the iterative process of hyperparameter optimization, the DNN model outputs detection results and determines whether a claim is fraudulent or not. [42]

1.5.2 ENSEMBLE MODELS

Machine learning model built using a classifier can be used to estimate new test samples, but it is difficult to have high accuracy with a single built classifier. Also, it cannot be applied all datasets to solve different problems. These lead to the generation of ensemble learning.

Ensemble methods are widely used production models in the industry recently because of the combination of the stacked models which has the capability of learning the whole dataset well without overfitting. Ensemble methods based on decision tree is used in this research. The basic overview of Decision tree is as below

Decision Trees (DTs) are a one of the supervised learning algorithms used for classification and regression which is non-parametric in nature. They can handle both numerical and categorical data for computation and particularly does not need one-hot encoding like some of other algorithms. It constructs tree in top-down strategy Decision tree splits the data for prediction and uses either one of the below as a measure to predict the class. [14]

The entropy measure for the tree is assigned as one if the sample of data after split from whole dataset is completely homogeneous.

$$\text{Entropy (t)} = -\sum p(i/t) \log_2 p(i/t)$$

Gini index measures the divergences between the probability distributions of the target values.

$$\text{Gini Index} = 1 - \sum [p(i/t)]^2$$

Information gain is the difference between the entropy of the node before splitting (parent node) and after splitting (child node).

$$\text{Info Gain} = E(\text{parent node}) - E(\text{child node})$$

E-Entropy

Gain Ratio is calculated and used to determine the goodness of a split

$$\text{Gain Ratio} = \text{Information gain} / \text{Entropy}$$

Decision Tree algorithm has various types of implementations which can be used based on dataset

- 1.ID3
- 2.CART
- 3.C4.5

1.5.2.1 RANDOM FOREST (BAGGING)

Random Forest algorithm is an easy to use machine learning algorithm that produces a decent accuracy on the prediction even without hyper-parameter tuning. Because of its simplicity and for the fact that it can be used for both classification and regression, it is also one of the widely used algorithms for solving data analysis driven problems.

Random forests are the aggregation of decision tree predictors such that each decision tree depends on the values of a random vector sampled independently and with the same dissemination for all trees in the forest. [15]

$$C_{Brf}(x) = \text{most vote } \{C_b(x)\} [\text{number of trees}(B)]$$

$C_b(x)$ - Prediction class of bth random forest [16]

1.5.2.2 GRADIENT BOOSTING ALGORITHM

GBM is a gradient boosting framework that uses tree-based learning algorithms. It is designed to be distributed and efficient with advantages such as faster training speed and higher efficiency, lower memory usage, Better accuracy, Support of parallel and GPU learning, Capable of handling large-scale data. Since it tends to overfit quickly it has enhancements to tune and improve with parameters such as tree constraints, shrinkage, random sampling and penalized learning.

The objective of any algorithm is to minimize loss function and the loss function for GBM like linear regression is given as

$$\text{Predictions} = y_{pi}$$

$$\text{Loss} = \text{MSE} = \sum (y_i - y_{ip})^2$$

y_i - ith dependent value, y_{pi} - ith independent, α - Learning rate will be used for tuning.

1.5.2.3 XGBOOST (BOOSTING ALGORITHM)

XGBoost stands for “Extreme Gradient Boosting” which has gradient boosted tree with objective functions that consists of two parts: training loss and regularization term:

$$\text{obj}(\theta) = L(\theta) + \Omega(\theta)$$

where L - training loss function, Ω - regularization term.

The training loss measures how predictive our model is with respect to the training dataset used for modeling. A common choice of L is the mean squared error and logistic loss functions, which is given by [17]

$$1. \text{MSE- } L(\theta) = \sum (y_i - y^{\wedge}_i)^2$$

$$2. \text{Logistic- } L(\theta) = \sum [y_i \ln(1 + e^{-y^{\wedge}_i}) + (1 - y_i) \ln(1 + e^{y^{\wedge}_i})]$$

1.6 EVALUATION AND INTERPRETABILITY

1.6.1. METRICS EVALUATION USING CROSS-VALIDATION (10-FOLDS)

The problem with a model that predicts the label with the training data with high accuracy and fails to predict the real-world new data is called over-fitting. To avoid this, a part of data is taken out kept as test data and used later for evaluation. Rest of the data is split into training and validation set for modeling and used with a method called cross-validation [18]. It is a statistical method used to estimate the ability of machine learning models and it is a resampling procedure used to evaluate model in small sample. Cross-validation has a single parameter called k which refers to the number of independent samples split that can be made from the whole data. Hence, the procedure is called as **k-fold cross-validation**. When a specific value for k is selected, such as k=5 becoming 5-fold cross-validation which has 5 validation samples for 5 models. Since it is very easy to implement, easy to understand, and results in good bias-variance tradeoff at the end, this method is mostly used in applied machine learning to evaluate and pick a model for a given problem.

1.6.2 MONTE-CARLO SIMULATION

Monte Carlo cross validation (MCCV) is a simple and effective procedure. In k-fold CV, a sequence of train samples are required whereas in MCCV, training process uses the set r and the next random number of observations. The number of random observations is determined by $v = n - r - h$ where n is the total number of data points available. After training, the model predicts h values [19]. Model predictions cannot be easily done due to the intervention of random variables. The impact of risk and uncertainty in forecasting and machine learning models can be analyzed using this technique and this is also referred as probability simulation technique. The input features or random variables are modelled based on probability distributions such as normal, log normal, etc. Different iterations or simulations are run for generating paths and the outcome is divided by the number of iterations to achieve overall score of the model.

1.6.3 LEARNING CURVE

There are two important errors that need to be handled to build an accurate model, which are bias and variance. A learning curve shows the changes in training and validation error as the training set size increases to the maximum. It is a plot where the horizontal axis displays the number of training samples given to the network and the vertical axis displays the error [21]. Two curves are plotted, one with error on the train set and the other one with error on the test set. When we increase the train set, it will reduce the error in the test set ideally. It will let us decide on estimation if the capacity of the model to fit the data is enough for the classification error. The validation set error should never be expected to be significantly lower than training set error. If it is too high, then even more data to train will not help but we should consider changing the algorithm used for modeling. This is a good way of building a model while we can visualize and conclude on the model's ability to generalize and to have high accuracy.

1.7 MODEL INTERPRETABILITY

“Interpretability is the degree to which a human being can interpret the reason of a decision” -Miller [22]. In traditional statistics, we formulate and verify hypotheses by analyzing large datasets. There is a ground truth that correlation often does not lead to causation. So, interpretation of a model is a must to decide on building reliable models. Though machine learning models are adopted in solving problems, they mostly remain and are seen as black boxes. Understanding the reasons behind the black box model is important to believe the model output and to take actions based on prediction and deploy it to the production systems. XAI (explainable artificial intelligence) is the solution to the human-agent interaction problem [22]. There are frameworks such as LIME, LOCO, SHAPLEY, DALEX, etc., which are very useful in interpreting the machine learning models.

1.7.1 LIME

LIME is a novel explanation technique that explains by learning an interpretable model locally around the prediction and explains the individual predictions of any classifier clearly in an interpretable and faithful manner [23].

It is a model-agnostic framework which means it can be applied to any built model with any algorithm which breaks the black-box model with high accuracy predictions. This technique analyses deeply to understand how it affects the prediction and understand the model by changing the data samples to input and understand completely to give explanation. A list of explanations which depicts the contribution of each attribute to the prediction of a data sample is the output of LIME explainer. This allows us to determine which attribute changes will have highest influence on the prediction and clearly provides local interpretability on the model.

The formula used in LIME is given as follows:

$$\xi(x) = \arg g \in G L(f, g, \pi x) + \Omega(g) \quad (1) \quad [23]$$

A variety of explanation families G , fidelity functions L , and complexity measures Ω can be calculated from the above formula. The theme of LIME models is on linear models as clear explanations about the prediction and calculating the search of importance using perturbations.

1.7.2 SHAPLEY

Scott M Lundberg from the University of Washington proposed a technique called SHAP values which is based on Shapley values, a technique used in game theory to find how each player in a collaborative game the contribution to success [24]. Shapley Feature IMPortance (SFIMP) measure allows to visualize and interpret the model and the importance of each feature to the model performance. The goal of this technique is to distribute the performance difference fairly among the individual features when all features are used and when all features are ignored as well to see the prime importance. This is a like

significant variables in logistic regression, where we can determine the influence of each feature by looking at the value of its coefficient.

SHAP values offer two important advantages

1. Any tree-based model can be calculated instead of restricting it only to simple linear models. Hence, we can build complex, non-linear with interpretation which leads us to build more accurate models.
2. Each data point will have its own SHAP values.

1.7.3 DALEX

DALEX methodology is model-agnostic which means it works for any predictive model and returns a numeric score for any classification and regression models. The explainers built using this method cannot be based on model parameters nor model structure in order to achieve a truly model-agnostic solution. The only assumption here is that the predict function can be called for any selected data points and this function is wrapped with the model built and the validation dataset which serves as a collective interface for a model. There are two separate functions included in better understanding of global structure of a model (a.k.a. model explainers) and for better understanding of a local structure of a model (a.k.a. prediction explainers) [25]. These functions explain a single feature of a model and hence called as explainers. Results from each explainer are numerical summaries in a tabular format which may be summarized with generic plot function. The plot function in DALEX package can interpret any number of models and help in visualizing all models in a single chart for cross evaluation.

1.8 RESTFUL API FOR PREDICTION

A RESTful API breaks down a transaction request to create a series of small modules. Each module present in API addresses a part of the transaction and provides developers with a lot of flexibility. REpresentational State Transfer (REST) architecture allows resources to be accessed using HTTP request methods. Each resource is located at an endpoint and identified by a URL which includes special variables in the form of path parameters and a query string. A path parameter is a part of the URL that specifies a resource and is denoted by a noun within braces ({...}) inside the endpoint [26]. The query string is a list of key-value pairs at the end of a URL and it is used to return structured representation. Resources within

the API are represented in JavaScript Object Notation (JSON) format and are referred to as objects. The user should specify how to represent an object when called using HTTP requests. HTTP request has methods such as GET, POST, DELETE, PUT. Common HTTP request methods are GET and POST. The table below explains the action by the HTTP requests

HTTP Method	URI	Action
GET	http://[hostname]/todo/api/v1.0/tasks	Retrieve list of tasks
GET	http://[hostname]/todo/api/v1.0/tasks/[task_id]	Retrieve a task
POST	http://[hostname]/todo/api/v1.0/tasks	Create a new task
PUT	http://[hostname]/todo/api/v1.0/tasks/[task_id]	Update an existing task
DELETE	http://[hostname]/todo/api/v1.0/tasks/[task_id]	Delete a task

Table 1 HTTP request methods [27]

1.8.1 API USING PLUMBER IN R

The **plumber** R package was developed by Trestle Technology, LLC in 2017 which allows developers to serve users with existing R code as a service on the web. It translates the annotations like @get, @post, etc., that we place in functions into an HTTP API that can be called from other machines on network. It can be deployed on cloud and can be used by public internet as well. The object developed must be serialized into some format that the client can understand to send a response from R server to an API client. JavaScript Object Notation (JSON) is the most commonly used object by web APIs. Error handling can be configured in the R function to capture errors in the input/operation. A simple function in R is given as [28]

```
#' Return "hello world"
#' @get /hello
function(){
  "hello world"
}
```

1.8.2 API USING FLASK IN PYTHON

Flask is a micro web framework developed in Python by Armin Ronacher in 2010. It is called as microframework because it does not need tools or libraries. It does not have database abstraction layer, form validation, or any other components like other frameworks where existing third-party libraries provide those common functions. Flask has inbuilt boilerplate code for getting a simple app up and running. A sample python code developed using flask is given as [29]

```
from flask import Flask
app = Flask(__name__)
@app.route('/')
def hello_world():
    return 'Hello, World!'
```

1.9 RESEARCH OBJECTIVE

The objective of the research is to examine different techniques to build a good classifier using the data. It also involves in investigating the potential of machine learning classifiers on how successfully it classifies the positive and negative classes based on the engineered features from the overall features. This research builds ensemble models for the classification which are called as black-box models and it is also looks to make it interpretable using packages like LIME. The comparison of ensemble models is made after the build and compare with the traditional models with learning curves on how generalized the model is. The recent evolution of ensemble models in the industry is the motivation for the experimenting on ensemble models. Also, A simple prediction API is built using plumber package in R and tested in postman which interacts with built API.

1.10 RESEARCH QUESTION

Insurance fraud claims is the major issue in the industry and there are a smaller number of researches specifically on Automobile insurance fraud claim than healthcare fraud claims. This research works on auto insurance fraud claims prediction and focuses on answering below questions

- 1.How far the sampling techniques and feature engineering increases the data quality?
- 2.Can ensemble models predict fraud claims than traditional models?
- 3.How influencing is the model interpretability/explain ability for a black-box model?
- 4.Can a RESTful API be built for fraud claims prediction?

2 LITERATURE REVIEW

Maria Fernanda, et al., researched about the behavior of different oversampling techniques through different classifiers and evaluation metrics. The techniques are Random oversampling, SOMO and SMOTE. A real data from a Colombian insurance company was used in the research in predicting fraudulent claims for its compulsory auto product. They concluded from the research and clearly demonstrated the advantages of using oversampling for imbalance circumstances but also the importance of comparing different evaluation metrics and classifiers to obtain accurate appropriate conclusions and comparable results [7]. This result was very helpful for deciding the method to use for imbalanced dataset

Nitesh V. Chawla, et al., formed a different technique in building classifiers from imbalanced datasets. A dataset is called as imbalanced dataset if the classification classes are not approximately equally represented. The real-world data sets have small number fraud instances. It also stated that the cost of misclassifying a fraud instance as a normal instance is often much higher than the cost of the reverse error. Under-sampling of the majority (normal) class has been proposed as a good means of increasing the sensitivity of a classifier in this research in equal weightage to the minority class. This paper found that better classifier performance (in ROC space) can be achieved by combining over-sampling of the minority (abnormal) class and under-sampling of the majority (normal) class than only under-sampling the majority class. This paper also shows that better classifier performance (in ROC space) can be achieved by combining over-sampling technique on minority class and under-sampling technique on majority class than varying the loss ratios or prior knowledge about classes in Naive Bayes. They also experimented in over-sampling the minority class which involves in creating synthetic minority class examples using SMOTE. These experiments were performed using C4.5 decision tree, Ripper and a Naive Bayes classifier. The evaluation strategy used in this investigation area under the Receiver Operating Characteristic curve (AUC) and the ROC convex hull strategy [8].

Sundara Kumar, et al., employed k Reverse Nearest Neighborhood and One Class support vector machine (OCSVM) which involves in rectifying the data imbalance problem with a hybrid approach. The demonstration of the proposed model with effectiveness was done by mining an Automobile Insurance

Fraud detection dataset and customer Credit Card Churn prediction dataset by following 10-fold cross validation method of testing using Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR), Probabilistic Neural Network (PNN), Group Method of Data Handling (GMDH), Multi-Layer Perceptron (MLP). High sensitivity value of 90.74% and 91.89% on Insurance dataset was achieved by DT and SVM. High sensitivity of 91.2%, 87.7%, and 83.1% on Credit Card Churn Prediction dataset was achieved by DT, SVM and GMDH respectively. They found that there is statistically no significant difference between DT and SVM in Insurance Fraud detection dataset. DT is preferred more than SVM since it has rule based predictions. It was found that GMDH, SVM and LR are not statistically different and GMDH yielded very high AUC at ROC with churn prediction dataset. In a significant outcome of this research, DT had minimum number of rules (4) on Insurance dataset and 10 rules on churn prediction datasets. An automobile insurance fraud dataset and a credit card customer churn dataset were also analyzed with one-class SVM with under sampling technique. [33]

Carol Hargreaves, et al., mentioned that Automobile insurance fraud is a global problem and handling fraud manually has always been costly for insurance companies. Machine Learning can play a crucial role in fraud detection and can aid insurance companies to identify fraud. This paper also proposes to determine which variables are significant for fraud detection and to provide a framework for the insurance fraud detection using **Chi-Square** and independent samples t-test [32]

Muhammad Fahim Uddin, et al., mentioned about the importance of feature engineering that ML requires a certain number of features to train the model. One of the main challenges is to determine the right quantity and the type of such features with great quality from whole dataset. It is not unusual for the ML process to use whole features in dataset without computing the importance value of each. If that approach is followed, it makes the process vulnerable to overfit, predictive errors, bias, and poor generalization. Each feature in the dataset has either a unique importance value, redundant, or irrelevant value. However, the way to get better accuracy and fitting for ML is to identify the optimum set number of right feature set with the finest matching of the importance value [9].

Riya Roy, et al., concluded from the research with sample of more than 500 data and dividing it into training and testing data. They compared with the algorithms based on the observation decision tree and random forest algorithms have better performance than naïve Bayes. [30]

Yoshihiro Ando, et al, observed that Random Forest is more accurate in detecting normal instances, and Neural Network is for detecting fraud instances and presented an ensemble method - based on a combination of random forest and neural network [31]

Carol Hargreaves, et al., proposed based on their research to leverage data analytics solutions to the fullest, insurance companies should use simple data analytic techniques such as statistical significance testing, then profiling of fraudulent claims by which business rules may be derived, after which a framework can be built. [32]

Fortuny et al. showed that SVM was a more appropriate algorithm than naive Bayes in fraud classification [34]. Meanwhile, Bhattacharyya et al. found that RF and SVM yielded a better result than logistic regression [35]

Whitrow et al. compared the performance of 7 algorithms, including RF, logistic regression, SVM, naive Bayes, quadratic discriminant analysis, CART and k-nearest neighbors, in fraud detection. The experimental results showed that RF and SVM performed better than other algorithms in fraud detection [36].

A report in 2017 was published by Nilson which mentioned that financial losses in card fraud related incidents reached 22.8 billion dollars to stress the importance of research and techniques in fraud claims. By 2021, This problem is forecasted in the credit card fraud bill to have losses around 32.96 billion dollars. [37]

Sinayobye Janvier Omar, et al., researched on State-of-the-Art Machine Learning techniques in fraud claims and mentioned that frauds in these systems are considered as cyber-crime, causing huge amount of financial losses. There are often two main criticisms of data mining-based fraud detection research: the deficiency of public available real data in this domain to perform experiments on I.e. Obtaining appropriate access to financial data to perform research in this area is extremely difficult due to privacy and competitive reasons, and the lack of published well-researched methods and techniques. They also stated the “No free lunch theorem” statement that there is no single machine learning technique that can uniformly outperform other technique over all datasets. They are in common having the same strengths and/or weaknesses. Fraud detection techniques have been categorized and reviewed. However, it is

noticed that most fraud detection systems in all areas use supervised approach. They also mentioned the most commonly used machine learning techniques are Artificial Neural Networks (ANN), Decision tree, Support Vector Machines (SVM), Naive Bayes, Random forest and K-NN algorithms [38].

Marco Tulio Ribeiro, et al., explains different models for text and image classification with interpretability methods. They depicted the use of these techniques on various scenarios that needs trust for making decision in prediction, choosing between models, improving an untrustworthy classifier, and identifying why a classifier should not be trusted. [23]

Leila, et al., studies the techniques for fraud detection in automobile insurance proposed by scientific studies from 1997 to 2016 (almost all papers in this regard). According to their analysis, Classification of data mining application class was used by 63% of the studies and is regarded as the best class of detection in automobile insurance. Among Classification techniques, Logistic Model, Decision Tree, and Naive Bayes had the highest rate of usage. [43]

3 RESEARCH METHODOLOGY

3.1. DATA DESCRIPTION

The main aim of this research is to make comparison between different ensemble algorithm techniques and traditional algorithms through SMOTE techniques and evaluation metrics like ROC with real data. This research used open-source dataset from Oracle in compulsory auto insurance claims. Imbalance phenomenon is presented with only 5.98% of fraudulent cases from 15,420 data points. The dataset has 33 features that describe the nature about the claims and the policy holder. These variables will be engineered to select few variables to classify between fraudulent and non-fraudulent claims.

The variable explanation chart is given below

Variable Name	Data Type	Variable Description
Month	factor	Month Of Accident that happened
weekOfMonth	integer	week Of Month that accident happened
DayOfWeek	factor	Day of the week that accident happened
Make	factor	Manufacturing company
AccidentArea	factor	Area where accident happened
DayOfWeekClaimed	factor	Claimed Day of week
MonthClaimed	factor	Month when the insurance was claimed
weekOfMonthClaimed	integer	week of month the amount was claimed
Sex	factor	Sex of the person who claimed
Maritalstatus	factor	Marital status of the claimed person
Age	integer	Age of the person
Fault	factor	Owner of the policy(Policy Holder or Third party)
Policytype	factor	Type of Insurance policy
VehicleCategory	factor	Category of the vehicle
VehiclePrice	factor	Price of the vehicle (In dollars)
FraudFound_P	integer	Label of the data(Fraud/Not Fraud)
PolicyNumber	integer	Number of the policy
RepNumber	integer	Repair number
Deductible	integer	Amount that can be Deductible
DriverRating	integer	Rating for the driver
Days_Policy_Accident	factor	Days Policy Accident
Days_Policy_Claim	factor	Days Policy Claimed
PastNumberOfClaims	factor	Total Number of claims in the past
AgeOfVehicle	factor	Age of the vehicle
AgeOfPolicyHolder	factor	Age of the Policy Holder
PoliceReportFiled	factor	Policy Report Filed date
witnessPresent	factor	witness Present or not?
AgentType	factor	Type of the agent
NumberOfSuppliments	factor	Number of Supplemnets done
AddressChange_Claim	factor	Address Change Claim
NumberOfCars	factor	Number of cars
Year	integer	Insurance formed Year
BasePolicy	factor	Insurance's Base Policy

Figure 1 Feature explanation

3.2 METHODOLOGY

The experimental research is based on the sampled dataset using a technique called SMOTE. we will use the sampled data through different classifiers such as Support Vector Machines, Decision Tree Classifier, Logistic regression and ensemble algorithms such as Random Forest, Gradient Boosting Machine and XGBoost algorithm. With distinct evaluation metrics: Accuracy, F1, Precision and Recall score, and ROC curve analysis. Various classifiers and evaluation metrics are used in this research to have robust results independent from the algorithm or metric used. Every experiment is executed in 5 folds using cross validation procedure. Then LIME model is built to see the explanation of the predicted result and to see variable importance. Then a REST API is built using plumber package in R and tested in postman for results of prediction.

3.2.1 H2O FRAMEWORK

H2O is an open source platform that allows us to build and deploy machine learning models in a scalable manner by processing big data with its in-memory, distributed, fast, highly efficient framework capable with maintenance in production systems of an enterprise environment. It is written in Java framework. Like some of NOSQL systems, H2O uses a Distributed Key/Value store to read and refer data, models, objects, etc., across all nodes and machines. The machine learning algorithms are executed on top of H2O's distributed Map/Reduce framework for processing big data. It processes quickly since it has parallel data loading feature and the data is fed through the clusters available, compressed and stored in memory in a columnar format. Data parser is very important for processing data and a built-in intelligence H2O's data parser is available to predict the schema of the input data and supports data ingestion from multiple sources in various formats [39].

3.2.2 DATA VISUALIZATION

Data Visualization involves in visualizing the influence of variables corresponding to the target variable to see how it makes difference in modeling and to get true understanding of the domain knowledge on the problem. R package called ggplot2 was used to explore the data visually and it follows grammar of graphics and easy to use. This exploratory data analysis gives us extensive knowledge on the insurance domain specifically on fraud claim problem.

A sample exploratory figure is given below which shows the distribution of our target variable

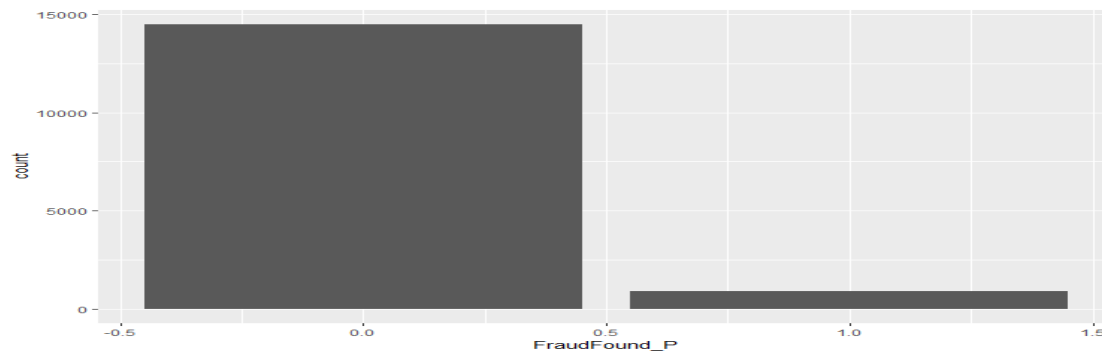


Figure 2 Distribution of label

3.2.3 FEATURE ENGINEERING

Feature Engineering is the most important section in building models. Feature engineering is possible only when we have domain knowledge and it helps us in building/understanding features that can be filtered out for the modeling. There are various packages like 'Boruta' in R and feature importance library in python. These kind of packages helps us in narrowing down the total number of the features that we can include for modeling. Also, we have traditional statistical feature filtering techniques like Chi-Square tests and Sample t-tests which works on the feature and target variable and gives the confidence p-value upon which we can decide to use the feature.

3.2.3.1 INDEPENDENT SAMPLE T-TEST AND CHI-SQUARE TEST:

Independent sample t-test is conducted for numerical attributes in the dataset. Chi-Square test is conducted for categorical attributes in the dataset in a statistical manner to see if there is a significant relationship between 2 categorical variables. Based on the test conducted, Features has been selected out of variables which has more than 5% confidence interval. The chi-square values calculated for individual variables corresponding to target variable is given as below

Variable	P-value	Variable	P-value
Month	0.001705	PastNumberOfClaims	1.43E-11
Make	2.20E-06	Age of Vehicle	0.002613
Accident Area	4.06E-05	AgeOfPolicyHolder	6.15E-05
Month Claimed	3.00E-05	PoliceReportFiled	0.05951
Sex	0.0002399	AgentType	0.006597
Fault	2.20E-16	NumberOfSuppliments	0.0004114
Policy Type	2.20E-16	AddressChange_Claim	2.20E-16
Vehicle Category	2.20E-16	Year	0.008321
Vehicle Price	2.98E-13	BasePolicy	2.20E-16
Deductible	1.30E-15	WeekOfMonth	0.6541
Days_Policy_Accident	0.02084	DayOfWeek	0.1185
Marital Status	0.798	DayOfWeekClaimed	0.6405
Driver Rating	0.6482	Days_Policy_Claim	0.1807
Witness Present	0.439	NumberOfCars	0.6597

Table 2 P-value from Chi-Square

3.2.4 DATA PRE-PROCESSING AND DATA PREPARATION:

Data pre-processing is the essential step before modeling any data. A clean data is more powerful than the hyperparameter tuning of any machine learning algorithm. Our data has lot of categorical variables, hence the prominent method called one-hot encoding is done to convert the variables from categorical variables to binary variables which is powerful for modeling and to avoid bias in the model. Data doesn't have much amount of missing values; hence we don't have to care about it. Ideally if the data has only 3-5 % of missing values, we can ignore that and process the rest of the data. It is essential to convert the target variable from integer to factor before modeling in order to train the model to understand that we need to predict the binary outcome. Since our data isn't much complicated, our pre-processing steps are

minimal. Before feeding the data into some algorithm, we must fix the imbalance in the data. There is a package in R called “DmWR” which has SMOTE function with it which can be used to balance our data before modeling.

3.2.4.1 DATA SPLITTING

Though we have learning curve as metrics to decide the optimal split of the data for modeling, we must specify the used data for modeling overall. When building a model, we need to have different splits in data so that we can validate the model without having any bias. The ideal way to split the data is into three sets namely training set (70%), validation set (15%) and test set (15%). The training set will be used to build the model and validation set will be used to validate the model by having different folds combined from training set using cross-validation technique. Finally, test set will be used like the real-world unseen data for the model for evaluation purpose.

3.3 AUTO ML

3.3.1 H2O AUTOML FRAMEWORK

H2O’s Auto ML is a big leap in AI world with open-source platform which takes us forward and makes things efficient by automating the machine learning workflow with automatic training and tuning of different machine learning algorithms and building models within the time specified by the user. It is the best in producing Stacked Ensemble model which is based on all previously trained models which stacks up the best model in learning the data without overfitting and comes up as a top performing models in the Auto ML Leaderboard. The top performing model can be used as a benchmark model for the task and tuning of hyperparameters can be done to surpass the leaderboard model from H2O. With H2O, we don’t need to resample the data to have balance in class since it has a parameter “balance-class” which will take care of that when it is set to “TRUE”. Auto ML is the big leap in the AI world which enables all engineers to leap forward to find the optimal solution.

3.3.2 RAPIDMINER AUTO ML FRAMEWORK

Auto Model was recently introduced in RapidMiner Studio to compete with the changing data science environment which accelerates the process of building and validating models by creating a process. This process can be modified or tuned well to improve with domain knowledge and deployed into production which justifies that there are no black-boxes anymore. Prediction, Clustering and Outliers are the problems addressed by Auto Model in RapidMiner currently. In case of Prediction category, classification and regression problems are solved. Evaluation of data is done using Auto Model which provides relevant models for solving the problem. It also helps in comparing the results for the built models, once the computations are done. Auto Model also helps in understanding the results even for Deep learning models which breaks the black-box model issue in data science community. Auto Model appears as a view in RapidMiner Studio which appears next to the Design view and the Results view [39]. This was also used to see and set the benchmark for modeling. Like H2O framework, parallel computing is available with machine learning algorithms such as Naive Bayes, Deep Learning, Random Forest, Gradient Boosting Machine, Logistic Regression, Generalized Linear Models and Decision Tree for modeling. Once the data is processed and models are built, it gives the different metrics and overall feature importance as well.

3.4 TRADITIONAL & ENSEMBLE MACHINE LEARNING ALGORITHMS

The execution of this experimental research is based on the comparison of machine learning algorithms which are used with balanced data with SMOTE technique. We execute through different classifiers: Decision trees, Logistic regression, Support vector machine, Random Forest, Gradient Boosting Machine and XGBoost with distinct evaluation metrics such as Accuracy, F1 score, Precision, Recall and ROC curve analysis.

Robust results are achieved with the use of many algorithms and evaluation metrics used to see and compare to come up with final model to deploy. Every experiment is repeated 5 times and consists of a 5-fold cross validation procedure. Also, Interpretability of the prediction is checked with LIME framework.

3.5 EVALUATION METRICS

3.5.1 ACCURACY PARADOX

“If you don’t know anything about Machine Learning, you should definitely know Accuracy Paradox” - Akinkunle Allen

Accuracy paradox is trusting accuracy as the only measure for the imbalanced data which leads to wrong conclusion. Classification models are very sensitive to class-imbalance in data and performs very poorly for model selection. Final decision in model selection should consider a combination of different metrics such as AUC, Precision, recall instead of relying on only one measure like accuracy which are calculated from confusion matrix in dealing with class-imbalance data. To report both the obtained performance measure values and the degree of class-imbalance in the data to minimize imbalanced-biased performance estimates is done when dealing with class-imbalance data. [6]

3.5.2 ACCURACY

Confusion matrix is used in order to validate performance of the algorithms. The main goal of confusion matrix is to compare predicted values and the actual ones as given in Table 1.

	Predicted:N	Predicted:P
Actual:N	TN	FP
Actual:P	FN	TP

Table 3 Confusion Matrix

From confusion matrix, we have True negatives (TN), True positives (TP), False positives (FP) and False negatives (FN). The values present in the diagonal represents those individuals whose classification are classified correct are TN and TP. The values present in the diagonal represents those individuals whose classification are classified wrong are FN and FP. Based on this, the predictive accuracy is defined as a performance evaluation measure and calculated by

$$P.A = \frac{(TP + TN)}{TP + FP + TN + FN}$$

It is always good to use different evaluation metrics when imbalance is presented in dataset than measuring the predictive accuracy measure.

3.5.3 ROC CURVE

Receiver operating characteristic (ROC) curve analysis is majorly used as a metric to measure the model reliability with AUC (Area Under Curve) score applied for dichotomic response variable and applied on imbalanced datasets. A range of tradeoffs between true positive and false positive error rates are done by summarizing the built model performance.

Defining TP as the true positive classified examples, TN as the true negative, FP as the false positive and FN as the false negative instances; false positive and true positive rates are expressed as follows

$$\text{False positive rate} = \frac{FP}{TN + FP}$$

$$\text{True positive rate} = \frac{TP}{TP + FN}$$

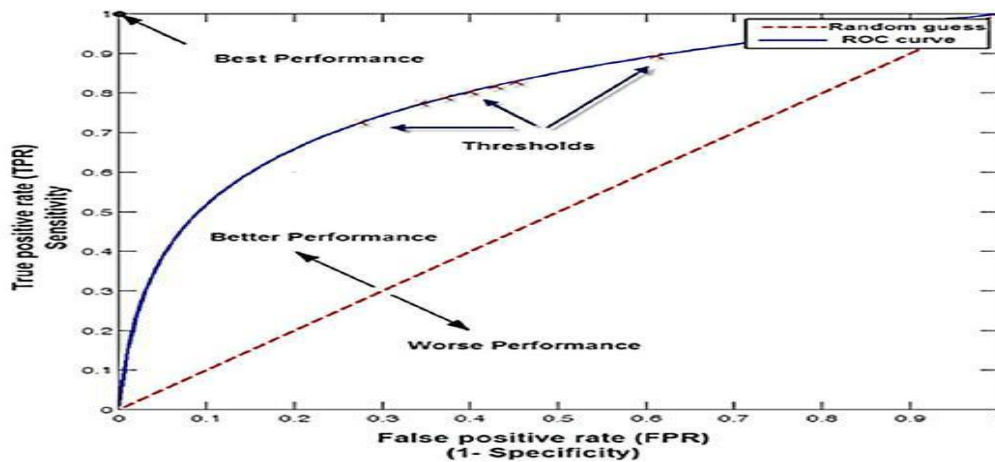


Figure 3 Receiver Operating Characteristics

ROC curve is obtained by plotting false positive rates (x-axis) versus true positives rates (y-axis), thus, the point (0,100) represents the ideal scenario in terms of misclassification errors which depicts all positive examples are classified correctly, and no negative examples are misclassified as positive.

Area under the ROC curve (AUC) consolidates the performance of a classification model into a single score range from 0 to 1 and not only allows comparing different ROC curves. It is very reliable in metrics measure than for a pure random classification model. If the AUC values are equal to 0.5, it is a good classifier and a good classifier always should reach an AUC larger than 0.5.

3.5.3 PRECISION AND RECALL

The precision-recall (PR) curve is good at explaining the tradeoff between precision which corresponds to false positive rate and recall which corresponds to false negative rate for different threshold values. A high AUC value represents both high recall which relates to a low false negative rate and high precision which relates to a low false positive rate.

Precision is defined as the total number of true positives (TP) from confusion matrix over the number of true positives (TP) plus the number of false positives (FP) from confusion matrix.

$$\text{precision} = \text{positive predictive value} = \text{TP} / (\text{TP} + \text{FP})$$

Recall (R) is defined as the number of true positives (TP) over the number of true positives plus the number of false negatives (FN).

$$\text{recall} = \text{sensitivity} = \text{TP} / (\text{TP} + \text{FN})$$

3.5.4 F1 SCORE

A good model should try to increase the recall value by increasing the true positive for the minority class and increase the precision value as well. This will be very hard when we are working with imbalanced datasets. This goal is very difficult to achieve when building the model using different algorithms.

ROC curve represents the tradeoff between TP and FP values. F score seeks to represent the trade-off between Precision and Recall with different values of TP, FP and FN. This evaluation metric can be expressed as follows:

$$F - value = \frac{(1 + \beta^2) * Recall * Precision}{(\beta^2 * recall) + precision}$$

Precision and recall have same meaningfulness is common to assume. ($\beta=1$)

3.6 LIME

The interpretability is achieved in LIME by first building LIME explainer with the training data used. The parameters like feature names, labels of the target, categorical variables are fed for building the LIME explainer. Then, we can pick any observation in the validation set for which explanation is required and get the explanation with initially built machine learning model by mentioning the total number of features we want in the explanation.

3.7 PLUMBER API

REST API can be made with model, server and plumber files. The model file consists of chosen built model which hold training set and all the memory needed to predict a new data. Server file loads the model into the function and consists of annotations like @post which receives input values and uses predict function with loaded model to give the output of given input values as a probability value to the user.

4 DATA ANALYSIS/FINDINGS AND DISCUSSION

4.1 BORUTA FEATURE SECTION

The working of Boruta package starts with making copy of the dataset and shuffling all the values in each feature which are called as shadow features. Then it trains an algorithm such as a Random Forest Classifier on the whole dataset. It ensures the importance of variables through the Mean Decrease Accuracy or Mean Decrease Impurity for each of the attributes in data. The algorithms check for higher score and gives importance to that feature accordingly. This will be compared with maximum Z-score [number of standard deviations from the mean a data point] of shadow features than the best of the shadow features. After some iterations, a final hit will be displayed. Validation of the feature importance takes place in checking the random forest algorithm's feature importance by comparing with random shuffled copies of the dataset and it totally increases the robustness of the model providing the important features which is done by comparing the real and shadow feature using a binomial distribution. If a feature dint gets recorded for 15 iterations, it gets rejected.

The sample feature selection done by the package is plotted as below with all features

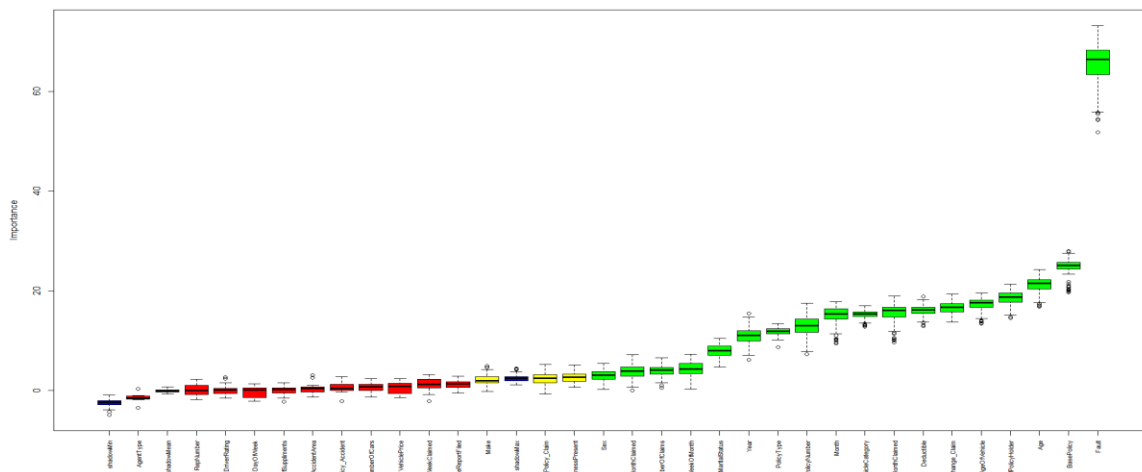


Figure 4 Boruta variable explanation boxplot

These selected features were compared with chi-square and independent statistics as well to filter final set of features. Domain understanding is also very important in selecting the final set of features for feeding it to the algorithms to build robust and accurate model.

Though the package selected 18 variables with more importance as below, the final selection should be critically analyzed with domain knowledge for feeding it to the algorithm.

The selected variables with importance values is given as below

Feature	Boruta Importance value
Month	14.996068
WeekOfMonth	4.289314
MonthClaimed	15.39794
WeekOfMonthClaimed	3.856009
Sex	2.952913
Marital Status	7.933663
Age	21.196082
Fault	65.164287
Policy Type	11.847488
Vehicle Category	15.224625
Policy Number	12.922261
Deductible	16.082393
PastNumberOfClaims	4.011812
AgeOfVehicle	17.252531
AgeOfPolicyHolder	18.482819
AddressChange_Claim	16.586144
Year	10.990497
Base Policy	24.764267

Table 4 Boruta Variable Importance

4.2 AUTO ML

Auto ML has become a wonderful evolution in the Data Science field which enable us to concentrate more on building accurate model than investing time in finding the best classifier out of hundreds of classifiers available in machine learning [40]. It analyzes and fits the best algorithm for the data and gives us the details including the hyperparameters used along with evaluation metrics.

The AUTO ML model built using H2O listed out 10 best models with the dataset and it is given as below

H2O Auto ML Model	AUC
GBM_grid_0_AutoML	0.953426
StackedEnsemble_AllModels_0_AutoML	0.953164
StackedEnsemble_BestOfFamily_0_AutoML	0.953123
GBM_grid_0_AutoML_20181229_211520_model_2	0.951868
GBM_grid_0_AutoML_20181229_211520_model_1	0.951321
GBM_grid_0_AutoML_20181229_211520_model_0	0.950323
GBM_grid_0_AutoML_20181229_211520_model_5	0.944166
GBM_grid_0_AutoML_20181229_211520_model_4	0.941972
XRT_0_AutoML_20181229_211520	0.940244
DRF_0_AutoML_20181229_211520	0.938419
GLM_grid_0_AutoML_20181229_211520_model_0	0.847305
DeepLearning_0_AutoML_20181229_211520	0.787446

Table 5 Auto ML Leader Board

Gradient Boosting Machine algorithm emerged as a top model in RapidMiner Auto Model framework with high AUC value 86% with sampled data.

The ROC comparison chart and evaluation metrics report from Rapid miner is given as below

ROC Comparison

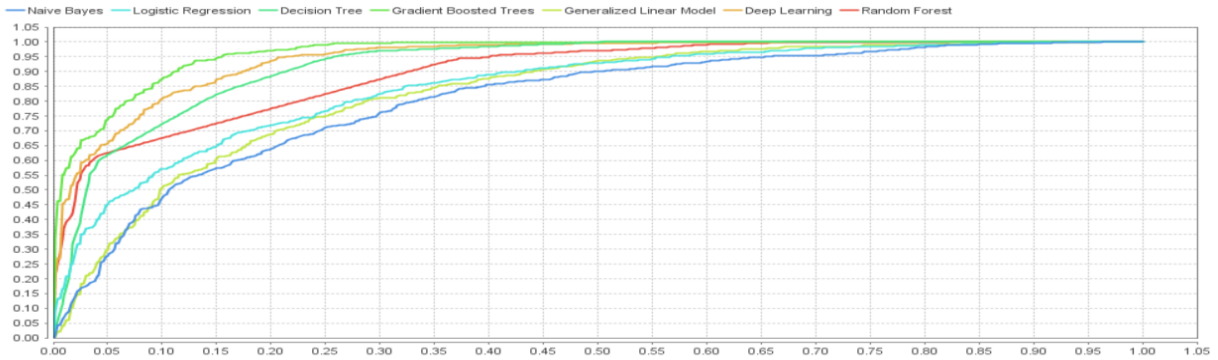


Figure 5 Receiver Operating Characteristics from RapidMiner Auto Model

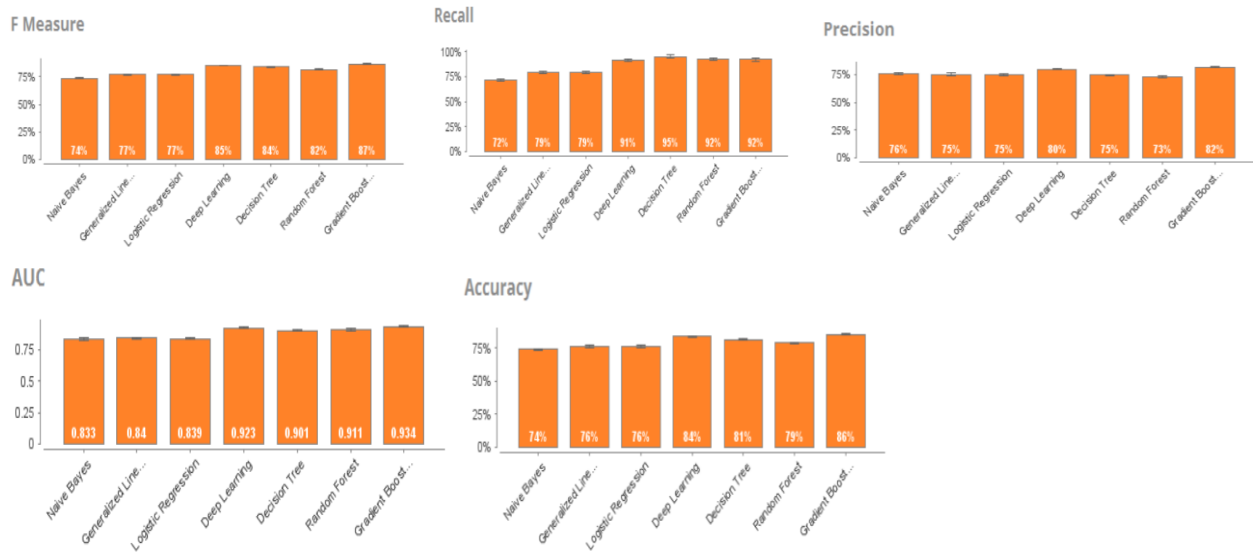


Figure 6 Metrics plot from RapidMiner Auto Model

4.3 ENSEMBLE MODELS

Ensemble models are so powerful since it learns by correcting itself through iterations. Learning curves were plotted for the tree-based models to see how the training and validation error using learning curve library in python with the sampled dataset. The learning curve for models are as below

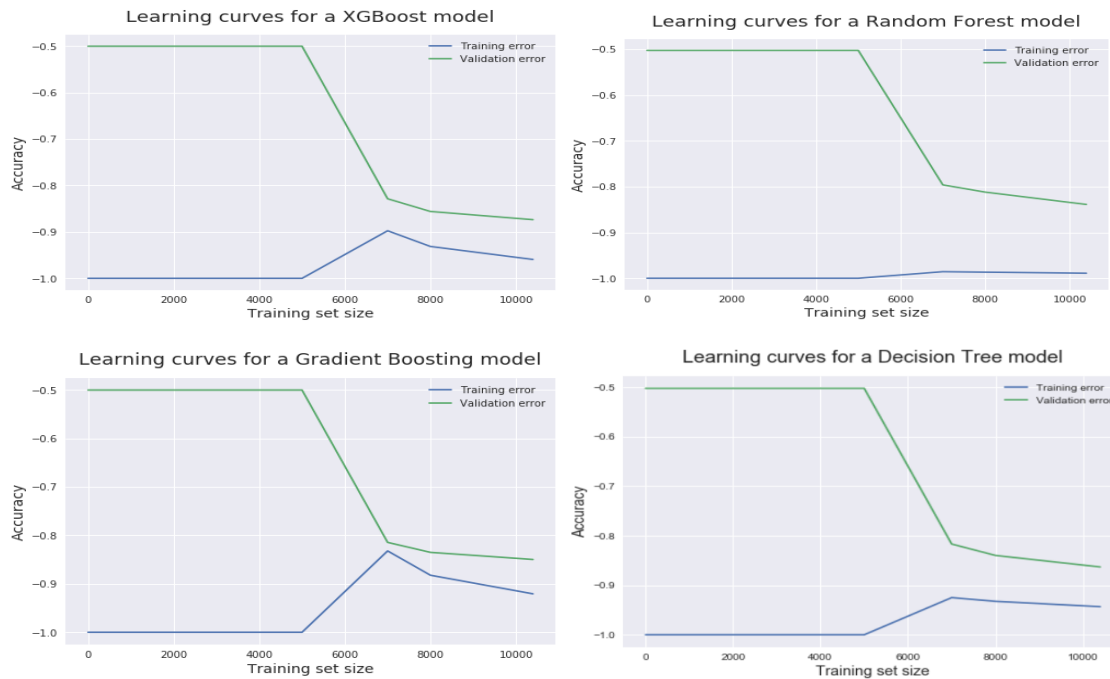


Figure 7 Learning curves for models

The convergence of XGBoost towards learning and validating seemed better than Gradient Boosting Machine, Decision Tree and Random Forest method.

4.4 OPTIMAL TREE DEPTH ANALYSIS FOR XGBoost:

Extreme Gradient Boosting Machine emerged as a best model for the dataset overall based on evaluation, the tree depth analysis was performed. Ideal number of trees for modeling is around 600 trees and maximum depth of the tree as 7.

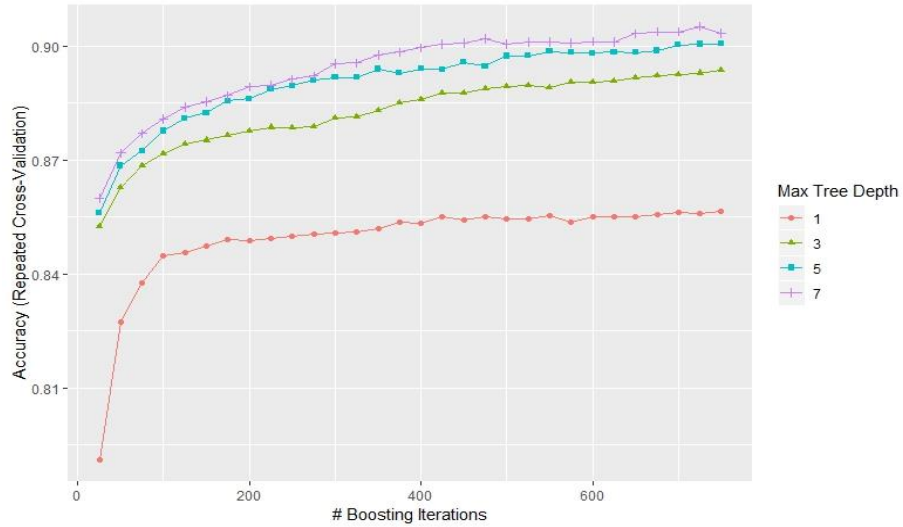


Figure 8 Optimal tree depth for XGBoost

4.5 RESULTS EXPLANATION

As we interpret from results table, boosting algorithms perform well on data than traditional algorithms. XGBoost performed well with high AUC score in concentrating positive class prediction with 86.8% AUC. Data was first split into train set and test set with ratio of 90% and 10% to train the classifier with high prediction capability and understand the nuance of the data.

The training data is sampled with SMOTE technique using ‘DmWr’ package in R and sampled to have classes balanced in the data. The sampled data was used in modeling as trainset, test set and validation set and split using H2O framework. Each model was trained using trainset and validated with 5 folds for validation and tested with test set. Again, the test set before sampling was used to test the model reliability in which XGBoost emerged with high AUC value.

The metrics was taken in accuracy, AUC and F1 to learn the reliability and understand the nuance in evaluating the model. Accuracy, AUC score and F1 score was high for the balanced data but when the imbalanced real data was used for testing, the model was able to capture True positive values and True negative values very well. The AUC score was good for XGBoost with 86.8% followed by GBM, Random Forest, GLM, ANN and SVM with 84.8%, 83.2%, 74.1%, 65.5% and 62% respectively.

F1 score for imbalanced data went down from balanced data F1 score since the recall value was low. But the model was good in predicting fraudulent claim and in non-fraudulent claim. Hence, AUC score was a good metric for this analysis. Hence the sampling techniques and feature engineering has increased the data quality before modeling.

Also, Ensemble models are better than traditional methods in predicting fraud claims and interpretability is influencing explain ability for a black-box models like ensemble models which gives us proper feature explanations to make decision based on that.

Algorithm	Set	Accuracy	AUC	F1
GLM	Training-H2O	0.764	0.834	0.764
	Test-H2O	0.763	0.843	0.781
	Imbalances test	0.939	0.741	0.223
ANN	Training-H2O	0.749	0.801	0.742
	Test-H2O	0.746	0.822	0.767
	Imbalanced test	0.939	0.655	0.176
SVM	Training	0.832	0.830	0.814
	Test	0.830	0.832	0.823
	Imbalanced test	0.879	0.62	0.243
Random Forest	Training-H2O	0.877	0.946	0.874
	Test-H2O	0.865	0.943	0.864
	Imbalanced test	0.944	0.832	0.319
GBM	Training-H2O	0.921	0.973	0.913
	Test-H2O	0.904	0.965	0.902
	Imbalanced test	0.945	0.848	0.360
XGBoost	Training-H2O	0.915	0.971	0.922
	Test-H2O	0.915	0.965	0.922
	Imbalanced test	0.948	0.868	0.428

Table 6 Metric table for built models

From the table, Training and Test in H2O are sampled data where the split was made using H2O framework and modeling was done. Imbalanced test is the real-world data split initially to test the model reliability.

4.6 INTERPRETABILITY OF BLACK BOX MODELS

LIME is a wonderful framework which explains the predictions so clearly with the influence of variables used. A sample LIME chart with our built model is as below

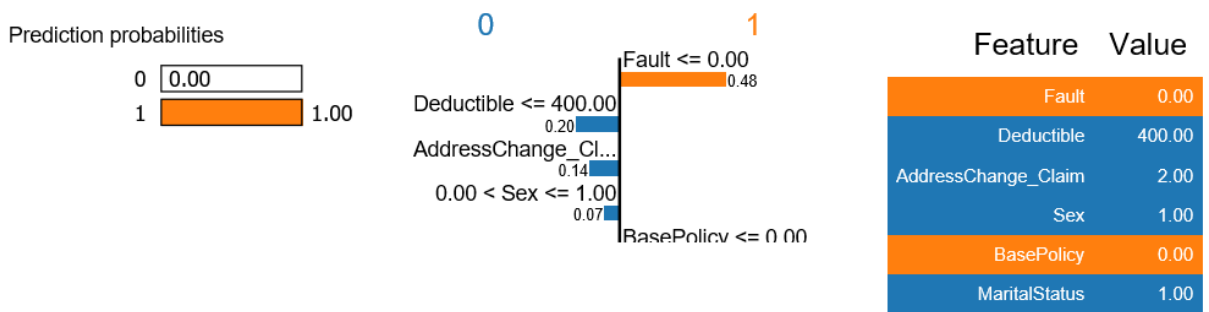


Figure 9 LIME prediction for XGBoost

4.7 REST API USING PLUMBER

API developed using plumber was tested in postman app with structured input. When input in JSON format is fed to the API, it is processed, and output is displayed in probability format. The screenshot is as below

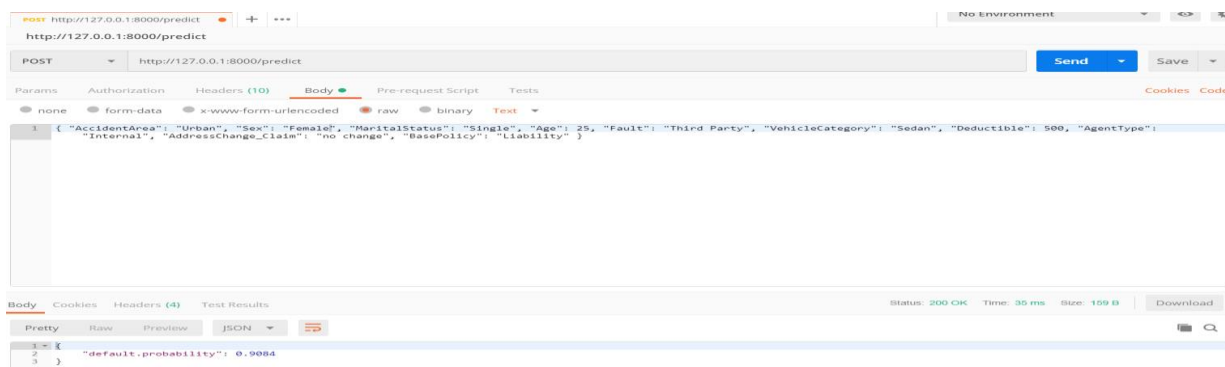


Figure 10 API output from postman app

5 CONCLUSIONS AND FUTURE WORK

Mining from imbalanced datasets is indeed a very important problem from both algorithmic and performance perspective [8], however, finding the correct methodology for doing it, brings important knowledge and solution for companies in making decisions such as detecting fraud claims in the insurance context. It is always good to have proper sampling techniques and feature engineering techniques as per the data to build a good classifier.

The use of predictive analytics in order to create preventive strategies in detecting fraud claims reaches as a crucial solution to the industry in recent days when we consider the loss that fraud incurs in auto insurance. In this research, we compared traditional machine learning techniques with ensemble machine learning algorithms across different metrics obtaining as a conclusion that an ensemble technique such as XGBoost technique had better performance results in most cases but also noted that recall value was low when calculating F1 metrics. The different evaluation metrics lead us to compare conclusions in each stage of the research.

In future, these results may be carried out and interpretability can be used in greater extent to understand the reason behind the prediction for better understanding. This shows that methodologies involving ensemble techniques produces better performance in classification tasks than traditional machine learning techniques on fraud detection contributing to the fight against the fraudulent activities and research in future should incorporate ensemble approach for a business problem. Also, some advanced techniques such as Natural Language Processing (NLP) can be done to collect more quality data from the description of the claim to improve the prediction level to the maximum.

5.1 REFLECTIONS

It was a wonderful journey in working with the dissertation. I thoroughly enjoyed and cherished my passion towards Artificial Intelligence with this research. I see a future myself as a Fraud/Risk Analyst because of the research I explored and witnessed the possibility of becoming one. The support and encouragement from my Supervisor were tremendous and it motivated me to learn contents and refer research papers to enhance my knowledge thoroughly.

My experience in master's program was very interesting and practical. The syllabus was intact and aligned with my interest since I wanted to become a Data Scientist though I wish there might have been electives for us to choose. I believe it will be implemented soon for upcoming students and make themselves to decide and choose on their own.

6 BIBLIOGRAPHY

- [1] S. Jordon, "Insurance fraud: 'its all over the place' and you should care about it officials say", Omaha World-Herald, 2016.
- [2] Y. Shi, C. Sun, Q. Li, L. Cui, H. Yu, C. Miao, "A fraud resilient medical insurance claim system", AAAI, pp. 4393-4394, 2016.
- [3] W. Raghupathi, V. Raghupathi, "Big data analytics in healthcare: promise and potential", Health Information Science and Systems, vol. 2, no. 1, pp. 3, 2014.
- [4] Legal dictionary <https://legaldictionary.net>
- [5] S. Tennyson, "Insurance experience and consumers' attitudes toward insurance fraud", Journal of Insurance Regulation, vol. 21, no. 2, pp. 35, 2002.
- [6] Josephine S Akosa, Oklahoma State University ,2017"Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data
- [7] María Fernanda Osorio Moreno, COMPARING THE PERFORMANCE OF OVERSAMPLING TECHNIQUES FOR IMBALANCED LEARNING IN INSURANCE FRAUD DETECTION, 2017
- [8] Nitesh V. Chawla, et al., SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research 16 (2002) 321–357
- [9] Muhammad Fahim Uddin, et al., Proposing Enhanced Feature Engineering and a Selection Model for Machine Learning Processes, Applied Sciences 2018
- [10] Mary L. McHugh, Lessons in biostatistics-The Chi-square test of independence, May 6, 2013
- [11] Miron Bartosz Kurska, Package 'Boruta' -Wrapper Algorithm for All Relevant Feature Selection, July 17, 2018
- [12] Jakkula, V. Tutorial on Support Vector Machine (SVM). Available online: <http://www.ccs.neu.edu/course/cs5100f11/resources/jakkula.pdf>, 15 March 2013
- [13] Hosmer DW Jr, et al., Applied Logistic Regression, Third Edition. New Jersey: John Wiley & Sons, 2013
- [14] Sonia Singh, et L., COMPARATIVE STUDY ID3, CART AND C4.5 DECISION TREE ALGORITHM: A SURVEY, 2014
- [15] Leo Breiman, RANDOM FORESTS, Statistics Department, January 2001
- [16] Trevor Hastie, et al., Elements of Statistical Learning, 2008
- [17] Official document for XGBOOST- <https://xgboost.readthedocs.io/en/latest/tutorials/model.html>
- [18] Official document https://scikit-learn.org/stable/modules/cross_validation.html

- [19] R.Fonseca-Delgado, et al., An assessment of ten-fold and Monte Carlo cross validations for time series forecasting, Sep 2013
- [20] J. Van Jaarsveld, F. Mostert, J. Mostert, "The claims handling process of liability insurance in South Africa", 2015.
- [21] Martin Thomas, Analysis and Optimization of Convolutional Neural Network Architectures, Jul 2017
- [22] Tim Miller, Explanation in Artificial Intelligence: Insights from the Social Sciences, Aug 2018
- [23] Marco Tulio Ribeiro, et al., "Why Should I Trust You?" Explaining the Predictions of Any Classifier, Aug 2016
- [24] Giuseppe Casalicchio, et al., Visualizing the Feature Importance for Black Box Models, Jul 2018
- [25] Przemyslaw Biecek, DALEX: Explainers for Complex Predictive Models in R, July 2018
- [26] Eric Gossett, et al., AFLOW-ML: A RESTful API for machine-learning predictions of materials properties, Sep 2018
- [27] Miguel Grinberg, <https://blog.miguelgrinberg.com/post/designing-a-restful-api-using-flask-restful>, 2013
- [28] Jeff Allen, Official Documentation-<https://www.rplumber.io/docs/>
- [29] Armin Ronacher, et al., Official documentation-
<https://media.readthedocs.org/pdf/flask/latest/flask.pdf>, Sep 2017
- [30] Detecting insurance claims fraud using machine learning techniques- IEEE-2017
- [31] Yoshihiro Ando, et al, Detecting Fraudulent Behavior Using Recurrent Neural Networks, 2015
- [32] Analytics for Insurance Fraud Detection: An Empirical Study –Jan 2016
- [33] Sundara kumar, et al., A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance, January 2015
- [34] E.J.d. Fortuny, et al., Corporate residence fraud detection, Aug 2014
- [35] S. Bhattacharyya, et al., Data mining for credit card fraud: a comparative study, 2011
- [36] C. Whitrow, et al., Transaction aggregation as a strategy for credit card fraud detection, 2009
- [37] Article- <https://www.kdnuggets.com/2018/08/introduction-fraud-detection-systems.html>
- [38] A State-of-the-Art Review of Machine Learning Techniques for Fraud Detection Research, 2018
- [39] Official Documentation - <https://docs.rapidminer.com/8.1/studio/auto-model/>
- [40] Manuel Fernandez-Delgado et al., Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? 2014
- [41] Jürgen Schmidhuber, Deep learning in neural networks: An overview, 2014

[42] Yibo Wang, et al., Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud, Nov 2017

[43] Leila Goleiji, et al., Survey of Detecting Fraud in Automobile Insurance Using Data Mining Techniques, Nov 2016