

**AUTOMOBILE INSURANCE CLAIM PREDICTION**

**INT 300 INTERNSHIP 2**

**PROJECT REPORT**

***Submitted by***

**SHREYA S – E0322030**

***In partial fulfilment for the award of the degree of***

**BACHELOR OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE AND ENGINEERING**

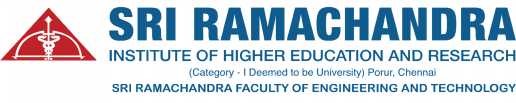
**(Artificial Intelligence and Data Analytics)**

**Sri Ramachandra Faculty of Engineering and Technology**

**Sri Ramachandra Institute of Higher Education and Research, Porur,**

**Chennai –** **600116**

**APRIL 2024**



**AUTOMOBILE INSURANCE CLAIM PREDICTION**

**INT 300 INTERNSHIP 2**

**PROJECT REPORT**

***Submitted by***

**SHREYA S – E0322030**

***In partial fulfilment for the award of the degree of***

**BACHELOR OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE AND ENGINEERING**

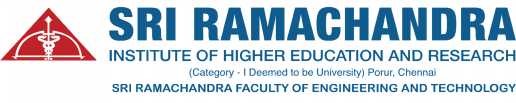
**(Artificial Intelligence and Data Analytics)**

**Sri Ramachandra Faculty of Engineering and Technology**

**Sri Ramachandra Institute of Higher Education and Research, Porur,**

**Chennai -600116**

**APRIL 2024**



**BONAFIDE CERTIFICATE**

Certified that this project report **“ AUTOMOBILE INSURANCE CLAIM PREDICTION ”** is the bonafide record of work done by **“SHREYA S - E0322030”** who carried out the internship work under my supervision.

**Signature of the Supervisor Signature of the HOD**

**Dr. Nirmala B Dr. Uma Satya Ranjan**

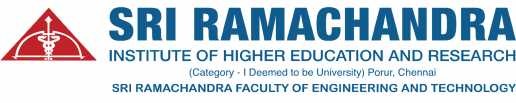
**Assistant Professor, Professor and Head,**

Department of Artificial Intelligence and Data Analytics Department of Artificial Intelligence and Data Analytics

Sri Ramachandra Faculty of Engineering and Technology, Sri Ramachandra Faculty of Engineering and

SRIHER, Porur, Chennai - 600116 Technology, SRIHER, Porur, Chennai-600 116.

**Evaluation Date:**



# ACKNOWLEDGEMENT

I express my sincere gratitude to our Dean of Sri Ramachandra Faculty of Engineering and Technology, Dr. T. Ragunathan for providing the required

facilities for this study.

[

I would also like to extend heartfelt thanks to our Head of the Department Dr. Uma Satya Rajan for helping us with additional support during this study.

[

I wish to thank my faculty supervisor(s), Dr. Nirmala, Department of Artificial Intelligence and Data Analytics, Sri Ramachandra faculty of Engineering and Technology and Mr. VEERAMANI KALYANASUNDARAM , REPORTING MANAGER ,LTIMINDTREE for extending help and encouragement throughout the project. Without his/her continuous guidance and persistent help, this project would not have been a success for me.

I am grateful to all the staff members of Sri Ramachandra Faculty of Engineering and Technology, my beloved parents and friends for extending the support, who helped us to overcome obstacles in the study.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO.** | **TITLE**    **ABSTRACT**  **LIST OF FIGURES** | **PAGE NO.**  6  7 |
| **1** | **INTRODUCTION** | 8 |
|  | * 1. Obstacles in Forecasting Automobile Insurance Claims | 9 |
|  | 1.2 Data Origins and Variables  1.3 Predictive Modeling Approaches  1.4 Advantages of Predicting Automobile Insurance Claims    1.5 Ethical Considerations and Equity    1.6 Future Prospects and Opportunities | 9  10  10  11  11 |
|  |  |  |
| **2** | **LITERATURE REVIEW** | 12 |
| **3** | **PROPOSED METHODOLOGY** | 14 |
| **4** | **IMPLEMENTATION** | 16 |
| **5** | **RESULTS AND DISCUSSIONS** | 18 |
|  | **APPENDICIES** | 20 |
|  | Appendix-1: Code Compiler | 20 |
|  | Appendix-2: Outputs | 28 |
|  | **REFERENCES** | 31 |
|  | **WORKLOG** | 32 |
|  | **CERTIFICATE OF COMPLETION** | 35 |

# ABSTRACT

This research delves into crafting a predictive framework aimed at gauging the likelihood of auto insurance claims. Given the burgeoning wealth of data within the insurance realm, there's a burgeoning interest in harnessing sophisticated analytics to gauge risk with precision. The proposed framework amalgamates machine learning methodologies to sift through historical data encompassing diverse variables like driver demographics, vehicle attributes, driving patterns, and prior claim histories. Techniques such as feature engineering and selection are deployed to pinpoint the pivotal factors influencing claim likelihood.

Diverse machine learning algorithms ranging from logistic regression to decision trees, random forests, and gradient boosting are scrutinized to pinpoint the most apt approach. The model is honed using a voluminous dataset of past insurance claims, with its efficacy gauged through cross-validation and assorted evaluation metrics encompassing accuracy, precision, recall, and F1score.

Moreover, interpretative techniques are applied to unravel the factors underpinning heightened claim likelihood, furnishing invaluable insights for insurance underwriting and risk mitigation. The resultant model strives to augment the precision and efficiency of auto insurance claim prognosis, thereby aiding insurance firms in more accurately pricing policies and navigating risks adeptly.

# LIST OF FIGURES

|  |  |  |
| --- | --- | --- |
| **FIG NO** | **DESCRIPTION** | **PAGE NO** |
| **1.** | **FRAUD REPORTED V/S COUNT (FIG 1.1)** | 30 |
| **2.** | **INCIDENT STATE V/S COUNT (FIG 1.2)** | 30 |
| **3.** | **POLICY STATE V/S COUNT (FIG 1.3)** | 30 |
| **4.** | **INCIDENT TYPE V/S COUNT (FIG 1.4)** | 31 |
| **5.** | **GENDER (FIG 1.5)** | 31 |
| **6.** | **RELATIONSHIP (FIG 1.6)** | 31 |
| **7.** | **ACCURACY (FIG 1.7)** | 32 |

# 

# CHAPTER 1

**INTRODUCTION**

Predicting automobile insurance claims stands as a pivotal element within the insurance sector, employing data analytics and machine learning methodologies to gauge the likelihood of forthcoming claims. With technological advancements and the burgeoning availability of data, insurance firms increasingly rely on predictive analytics to accurately evaluate risks and refine their business strategies.

When policyholders file for automobile insurance claims due to accidents, theft, vandalism, or other covered incidents, it necessitates the prediction of claim frequency and severity for insurers to determine appropriate premiums, manage reserves, and mitigate financial risks.The predictive process entails scrutinizing diverse factors contributing to the likelihood of incidents, encompassing the driver's demographics, driving history, vehicle specifications, geographical location, weather conditions, and historical claim trends.

Through analyzing past data patterns, insurers construct predictive models to anticipate future claims for varying policyholders. Employing predictive modelling techniques like regression analysis, decision trees, random forests, and neural networks aids in dissecting insurance claim data and pinpointing pertinent risk indicators. These models enable insurers to gauge policy-related risks, adjust premiums accordingly, and enact targeted risk mitigation measures.

The merits of automobile insurance claim prediction are multifaceted, ranging from refining underwriting accuracy and pricing strategies to combating fraudulent claims and enhancing customer service.

Nonetheless, challenges such as data integrity, model transparency, and regulatory compliance underscore the need for insurers to ensure access to quality data and adhere to regulatory standards.

In essence, automobile insurance claim prediction serves as a linchpin for insurers to effectively manage risks, boost operational efficiency, and elevate customer satisfaction amidst the competitive insurance landscape. By harnessing advanced analytics and machine learning, insurers glean insights into future claim patterns, empowering data-driven decision-making for sustained business success.

* 1. **Obstacles in Forecasting Automobile Insurance Claims:**

The task of predicting automobile insurance claims presents numerous challenges. Insurance firms grapple with copious amounts of data, spanning driver demographics, vehicle specifics, driving patterns, and historical claims records. Analyzing this data effectively demands advanced statistical methods, machine learning algorithms, and robust computational infrastructure. Moreover, factors like fraudulent claims, evolving regulations, and external economic variables further complicate the prediction process.

**1.2 Data Origins and Variables:**

The accuracy of predicting automobile insurance claims heavily relies on the availability and quality of data. Insurers typically gather data from diverse sources, including policyholder details, telematics devices, claims archives, and external data vendors. Variables such as age, gender, driving history, vehicle type, geographic location, driving behaviors, and past claims records commonly inform predictive models. Furthermore, advancements in IoT technology empower insurers to access real-time data on vehicle performance, driving habits, and environmental factors, bolstering the precision of predictive models.

**1.3 Predictive Modeling Approaches:**

Various predictive modeling techniques are employed in forecasting automobile insurance claims, ranging from conventional statistical methods to sophisticated machine learning algorithms. Linear regression, logistic regression, decision trees, random forests, and neural networks are some prevalent techniques. These models scrutinize historical data to unearth patterns and correlations among different variables, enabling insurers to anticipate the probability of future insurance claims. Additionally, ensemble methods and deep learning algorithms are increasingly being explored to enhance prediction accuracy and resilience.

**1.4 Advantages of Predicting Automobile Insurance Claims:**

Implementing robust systems for predicting automobile insurance claims offers manifold benefits to insurance companies, policyholders, and society at large. By accurately assessing risks, insurers can optimize pricing tactics, offer tailor-made insurance products, and elevate customer contentment. Furthermore, proactive risk management curtails the frequency and severity of insurance claims, resulting in reduced premiums for policyholders and augmented profitability for insurers. From a societal standpoint, efficient claim prediction bolsters the stability of the insurance market, fosters economic prosperity, and encourages safer driving practices.

**1.5 Ethical Considerations and Equity:**

While predicting automobile insurance claims holds immense promise, it also raises pertinent ethical concerns, particularly surrounding fairness and bias. Insurers must ensure that predictive models are developed and deployed transparently, accountably, and without bias. Techniques such as fairness-aware machine learning and bias mitigation strategies, alongside regulatory oversight, are pivotal in addressing these concerns and fostering equitable outcomes for all stakeholders.

**1.6 Future Prospects and Opportunities:**

As technology advances, the realm of predicting automobile insurance claims is ripe for further innovation. The integration of emerging technologies such as artificial intelligence, blockchain, and edge computing holds the potential to refine prediction accuracy, diminish administrative burdens, and streamline claims processing. Moreover, collaborative endeavors among insurers, researchers, and policymakers are imperative to tackle emerging challenges, including cybersecurity threats, data privacy issues, and regulatory compliance demands.

**CHAPTER 2**

**LITERATURE REVIEW**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.NO** | **YEAR** | **AUTHOR OR DEVELOPER** | **PROJECT** | **REMARKS** |
|  | 2019 | Jessica pesantez narvaez | Predicting motor insurance claims using telematics | This research compared the performances of logistic regression and XGBoost techniques to predict the existence of accident claims with a little number of data training, their results showed that logistic regression is an appropriate model given its interpretability and good predictive ability |
| **2.** | 2019 | Oskar Sucki | Churn prediction | It was found that random forests the best performing model (74% accuracies). The dataset had missing values in multiple fields. After looking at the distributions, it was decided to replace the missing variables with extra attributes that would indicate not having that information |
| 3. | 2018 | Muhammad Arief Fauzan | Accuracy of XGBoost for Predicting Claims | To compare the performance a set of techniques i.e., AdaBoost, Random Forest, Neural Network, with the performance of XGBoost. XGBoost gives better accuracies in terms of normalized Gini. |
| 4. | 2018 | G.Kowshalya, M.Nandhini. | Fraudulant Claims | To calculate insurance premium amount for different customers based on their personal and financial details, three classifiers were built to predict fraudulent claims and percentage of the premium amount. |
| 5. | 2017 | Tim Pijl | Applied knowledge discovery | The results show that dimensionality reduction is not necessarily needed for this problem and those simple techniques, such as a decision tree or random forest, outperform the more statistically advanced techniques, such as a support vector machine, also use small data set. |

**CHAPTER 3**

**PROPOSED METHODOLOGY**

**1. Data Collection and Cleaning:**

* Gather relevant data including policyholder information (age, gender, location), vehicle details (make, model, age), historical claims data (frequency, severity), and external factors (weather, traffic patterns).
* Clean the data to remove errors, missing values, and inconsistencies.

**2. Feature Engineering:**

* Extract meaningful features from the collected data. This might include:
* Policyholder features: Age, gender, occupation, marital status, credit score, driving history, etc.
* Vehicle features: Make, model, age, mileage, value, etc.
* Environmental features: Weather conditions, road type, traffic density, etc.
* Transform categorical variables into numerical representations using techniques like one-hot encoding or label encoding.
* Create new features if necessary, such as the age of the vehicle at the time of the claim, the ratio of past claims to years insured, etc.

**3. Exploratory Data Analysis (EDA):**

* Explore the relationships between different features and the target variable (i.e., whether a claim was made or not).
* Identify correlations and patterns in the data.
* Visualize the data to gain insights.

**4. Model Selection:**

* Choose appropriate machine learning models for prediction. Common models for insurance claim prediction include:
* Logistic Regression
* Decision Trees
* Random Forests
* Gradient Boosting Machines (e.g., XGBoost, LightGBM)
* Neural Networks
* Consider ensemble methods for improved performance.

**5. Model Training and Validation:**

* Split the data into training and validation sets.
* Train the selected models on the training data.
* Validate the models using the validation set and adjust hyperparameters as needed to improve performance.
* Evaluate models using appropriate metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC).

**6. Model Interpretation:**

* Interpret the trained models to understand which features are most important in predicting insurance claims.
* Understand how each feature contributes to the prediction.
* Identify any biases or confounding factors.

**7. Deployment and Monitoring:**

* Deploy the trained model into the insurance claim processing system.
* Continuously monitor the model's performance and retrain it periodically with new data.

**CHAPTER 4**

**IMPLEMENTATION**

1. **Import Libraries:**

Import necessary libraries such as pandas, numpy, seaborn, matplotlib, warnings, and required modules from scikit-learn.

1. **Load Data:**

Load the insurance claims data from a CSV file into a DataFrame named df.

1. **Data Exploration and Visualization:**

* Check the shape and info of the Data Frame.
* Check for missing values in the Data Frame.
* Visualize the distribution of the target variable fraud reported.
* Visualize the distribution of other categorical variables like incident state, policy state, incident type, insured sex, insured relationship, etc.
* Handle missing values by replacing '?' with and then impute missing values using the mode for categorical columns.
* Visualize the correlation matrix using a heatmap.

**4. Feature Selection:**

* Select the features to be used for modeling.
* Separate the features (x) and the target variable (y).

**5. Data Splitting:**

* Split the data into training and testing sets using train test split.
* Handle class imbalance using SMOTE (Synthetic Minority Over-sampling Technique).

**6. Model Building:**

* Define a stacking ensemble model using Stacking Classifier with base models including Decision Tree, KNN, and XGBoost, and Logistic Regression as the final estimator.
* Define individual models like Decision Tree, Random Forest, KNN, XGBoost, and the stacking ensemble.

**7. Model Evaluation:**

* Evaluate each model using cross-validation with repeated stratified k-fold.
* Calculate and print mean and standard deviation of accuracy scores for each model.
* Visualize the performance of each model using boxplots.

**8. Model Serialization:**

Save the trained stacking ensemble model (model) using pickle to a file named "insurance\_ml3.pkl".

**CHAPTER 5**

**RESULTS AND DISCUSSIONS**

Predicting automobile insurance claims is a complex task that involves analyzing various factors such as driver demographics, vehicle characteristics, historical claim data, and external factors like weather conditions and road infrastructure. Machine learning algorithms are often employed to analyze these factors and predict the likelihood of an insurance claim being filed.

In conclusion, automobile insurance claim prediction is valuable for insurance companies to assess risk accurately, set premiums effectively, and allocate resources efficiently. By leveraging advanced analytics and machine learning techniques, insurers can improve their underwriting processes, enhance customer service, and ultimately reduce costs associated with claims payouts. However, it's crucial to continuously refine and update these prediction models to adapt to changing circumstances and ensure their accuracy and reliability over time. Moreover, ethical considerations regarding data privacy and fairness in algorithmic decision-making must be carefully addressed to maintain trust and transparency in the insurance industry.

**RESULTS**

1. **Performance Metrics:**

Begin by presenting the performance metrics used to evaluate the predictive model. Common metrics include accuracy, precision, recall, F1 score, and area under the ROC curve (AUC). Provide the values for each metric obtained during model evaluation.

1. **Comparison with Baseline:**

Compare the performance of your predictive model with a baseline model or existing industry standards. Highlight any improvements in predictive accuracy or efficiency achieved by your model.

1. **Model Evaluation:**

Detail the performance of the predictive model across different evaluation datasets (training, validation, and test datasets). Discuss any variations in model performance across these datasets and identify potential areas for improvement.

1. **Feature Importance:**

Analyze the importance of features used in the predictive model. Identify which features contribute the most to accurate predictions and which ones have minimal impact.

1. **Confusion Matrix:**

Present the confusion matrix to illustrate the model's ability to correctly classify insurance claims into different categories (e.g., fraudulent vs. legitimate claims). Discuss any patterns or trends observed in the confusion matrix.

1. **Model Calibration:**

Assess the calibration of the predictive model by comparing predicted probabilities with actual outcomes. Use calibration plots or reliability diagrams to visualize the calibration performance.

1. **Cross-Validation Results:**

If applicable, discuss the results of cross-validation experiments to validate the robustness and generalizability of the predictive model.

**APPENDICES**

**APPENDIX-1:**

**CODE COMPILER**

import pandas as pd

import numpy as np

import warnings

warnings.filterwarnings('ignore')

import seaborn as sns

import matplotlib.pyplot as plt

pd.set\_option('display.max\_columns',50)

pd.set\_option('display.max\_rows',100)

df = pd.read\_csv(r"C:\Users\Shreya\Downloads\insurance\_claims.csv")

df

df.shape

df.info()

df.isna().sum()

print(df['fraud\_reported'].value\_counts())

sns.countplot(x='fraud\_reported',data=df)

df['incident\_state'].value\_counts()

sns.countplot(x='incident\_state',data=df, hue="fraud\_reported")

sns.countplot(x='policy\_state',data=df, hue='fraud\_reported')

plt.figure(figsize=(10,5))

sns.countplot(x='incident\_type',data=df, hue='fraud\_reported')

fig = plt.figure(figsize=(6,4))

ax=(df['insured\_sex'].value\_counts()\*100.0/len(df))\

.plot.pie(autopct='%.2f%%',labels =['Male','Female'],fontsize=12)

ax.set\_title('Gender %')

ax.set(ylabel='')

plt.show()

fig = plt.figure(figsize=(10,6))

ax = (df['insured\_relationship'].value\_counts()\*100.0/len(df))\

.plot.pie(autopct='%.1f%%',labels=['husband','wife','own child','unmarried','other relative','not in family'],

fontsize=12)

ax.set\_title('Realationship %')

# Hide y-axis label

ax.set(ylabel='')

plt.show()

df.head()

df.police\_report\_available.value\_counts()

df.property\_damage.value\_counts()

df.collision\_type.value\_counts()

df.replace('?', np.nan, inplace = True)

df.isna().sum()

from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()

df['insured\_sex'] = le.fit\_transform(df['insured\_sex'])

df['insured\_relationship'] = le.fit\_transform(df['insured\_relationship'])

df['incident\_type'] = le.fit\_transform(df['incident\_type'])

df['policy\_state'] = le.fit\_transform(df['policy\_state'])

df['insured\_education\_level'] = le.fit\_transform(df['insured\_education\_level'])

df['insured\_occupation'] = le.fit\_transform(df['insured\_occupation'])

df['insured\_hobbies'] = le.fit\_transform(df['insured\_hobbies'])

df['insured\_relationship'] = le.fit\_transform(df['insured\_relationship'])

df['collision\_type'] = le.fit\_transform(df['collision\_type'])

df['incident\_severity'] = le.fit\_transform(df['incident\_severity'])

df['authorities\_contacted'] = le.fit\_transform(df['authorities\_contacted'])

df['incident\_state'] = le.fit\_transform(df['incident\_state'])

df['incident\_city'] = le.fit\_transform(df['incident\_city'])

df['incident\_location'] = le.fit\_transform(df['incident\_location'])

df['property\_damage'] = le.fit\_transform(df['property\_damage'])

df['police\_report\_available'] = le.fit\_transform(df['police\_report\_available'])

df['auto\_make'] = le.fit\_transform(df['auto\_make'])

df['auto\_model'] = le.fit\_transform(df['auto\_model'])

df['fraud\_reported'] = le.fit\_transform(df['fraud\_reported'])

df.head(3)

import matplotlib.pyplot as plt

plt.figure(figsize = (20,15))

sns.heatmap(data = df.corr(),annot = True, fmt = '.2g', linewidth = 1)

plt.show()

corrs = df.corr()['fraud\_reported']

columns = corrs[corrs > .001].index

corrs = corrs.filter(columns)

corrs.sort\_values(ascending=False)

df.columns

df.head

feat=['vehicle\_claim','total\_claim\_amount','property\_claim','injury\_claim','umbrella\_limit',\

'number\_of\_vehicles\_involved','witnesses','bodily\_injuries','insured\_sex','policy\_state',

'insured\_relationship','months\_as\_customer']

x = df[feat]

y = df.fraud\_reported

X = df[feat]

y = df.fraud\_reported

from sklearn.model\_selection import train\_test\_split

x\_train, x\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size =0.3, random\_state = 42)

from imblearn.over\_sampling import SMOTE

from sklearn.model\_selection import train\_test\_split

x\_train, x\_test, y\_train, y\_test = train\_test\_split(x, y, random\_state = 1, stratify=y)

y\_train.value\_counts()

smt = SMOTE()

x\_train, y\_train = smt.fit\_resample(x\_train, y\_train)

np.bincount(y\_train)

from xgboost import XGBClassifier

from sklearn import model\_selection

from sklearn.linear\_model import LogisticRegression

from sklearn.linear\_model import LogisticRegressionCV

from sklearn.tree import DecisionTreeClassifier

from sklearn.neighbors import KNeighborsClassifier

from sklearn.ensemble import RandomForestClassifier

from sklearn.ensemble import StackingClassifier

from sklearn.model\_selection import cross\_val\_score

from sklearn.model\_selection import RepeatedStratifiedKFold

from numpy import mean

from numpy import std

#get a stacking ensemble of models

def get\_stacking():

# define the base models

level0 = list()

level0.append(('DT', DecisionTreeClassifier(max\_depth=10, random\_state=5)))

level0.append(('KNN', KNeighborsClassifier(5)))

level0.append(('XGB',XGBClassifier(objective= 'binary:logistic', use\_label\_encoder=False)))

level1 = LogisticRegression()

# define the stacking ensemble

model = StackingClassifier(estimators=level0, final\_estimator=level1, cv=10)

return model

# get a list of models to evaluate

def get\_models():

models = dict()

models['DT'] = DecisionTreeClassifier(max\_depth=10)

models['RF'] = RandomForestClassifier(n\_estimators=500)

models['KNN'] = KNeighborsClassifier(5)

#models['ADA'] = AdaboostClassifier(n\_estimators=500)

models['XGB'] = XGBClassifier(objective = 'binary:logistic', eval\_metric='logloss',use\_label\_encoder=False)

models['Stacking']= get\_stacking()

return models

#get the models to evaluate

models = get\_models()

#evaluate a give model using cross-validation

def evaluate\_model(model, x, y):

cv = RepeatedStratifiedKFold(n\_splits=10, n\_repeats=3, random\_state=5)

scores =cross\_val\_score(model, x, y, scoring='accuracy', cv=cv, n\_jobs=-1, error\_score='raise')

return scores

#evaluate the models and store results

results, names = list(), list()

for name, model in models.items():

scores = evaluate\_model(model, x\_train, y\_train)

results.append(scores)

names.append(name)

print('>%s %.2f (%.2f)' % (name,mean(scores), std(scores)))

plt.boxplot(results, labels=names, showmeans=True)

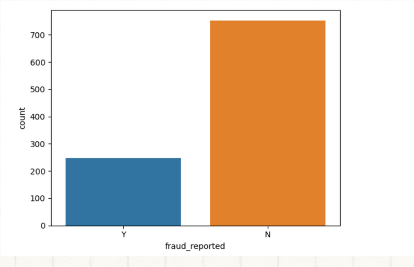
plt.show()

import pickle

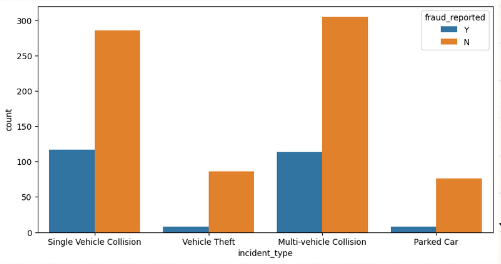
#save trained model to file

pickle.dump(model, open("insurance\_ml3.pkl", "wb"))

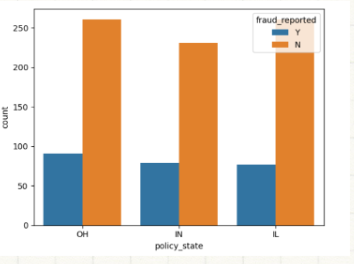
**APPENDIX-2: OUTPUT**

****

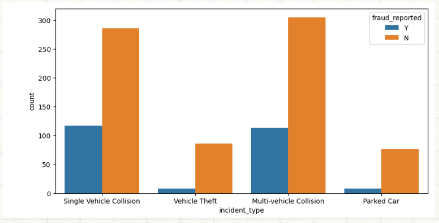
**SS 1.1**

****

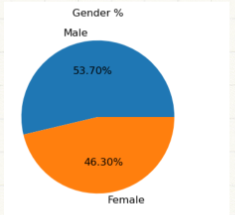
**SS 1.2**

****

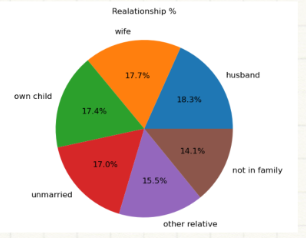
**SS 1.3**

****

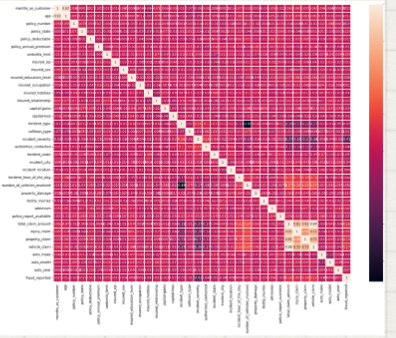
**SS 1.4**

****

**SS 1.4**

****

**SS 1.5**

****

**SS 1.6**

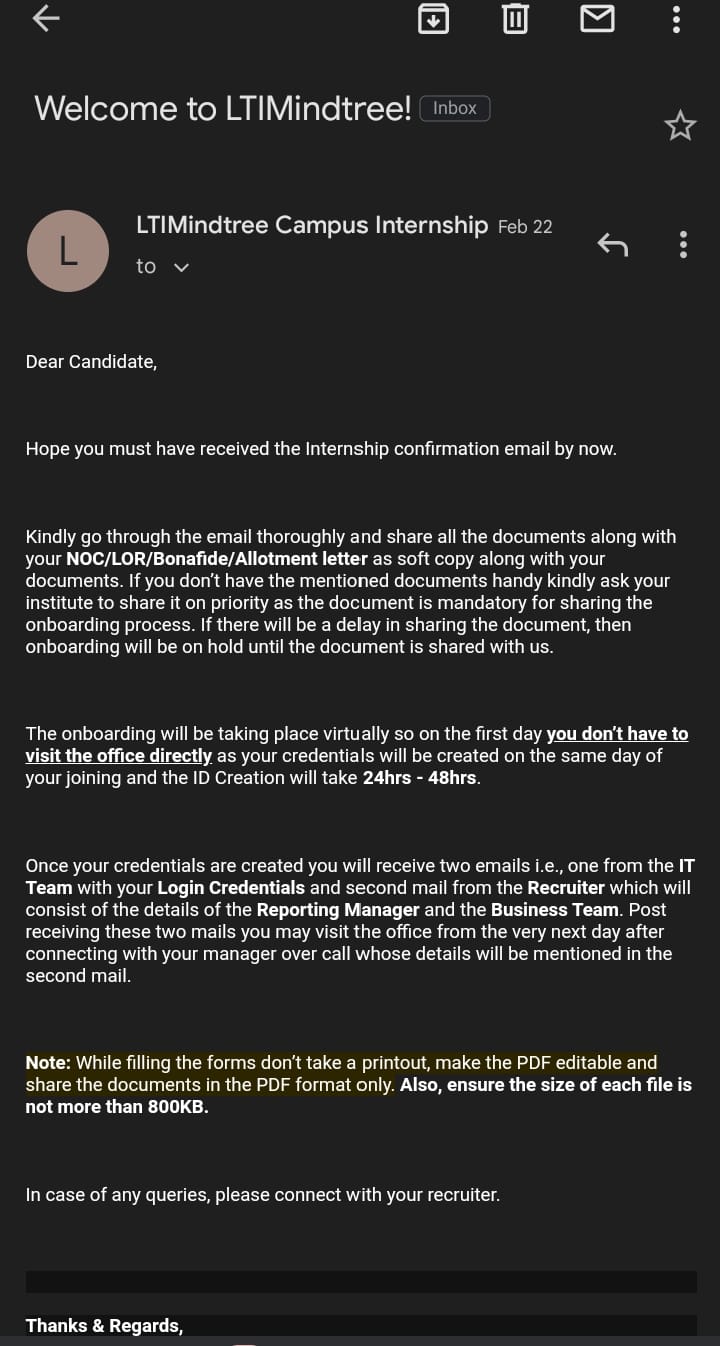
**REFERENCE**

1. Hanafy, M; Ming, R. Machine Learning Approaches for Auto Insurance Big Data. Risks 2021, 9(2) 42.
2. Hossin, M.; Sulaiman, M.N. A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process, 2015, 5(2), 01–11.
3. Fern´andez, A.; Garc´ıa, S.; Galar, M; Prati, R.C.;Krawczyk, B. Learning from Imbalanced Data Sets, 1st ed.; Springer Nature Switzerland AG, 2018.
4. Smith, K.A.; Willis, R.J.; Brooks, M. An analysis of customer retention and insurance claim patterns using data mining: a case study. Journal of the Operational Research Society, 2000, 51:5, 532-541.
5. Pesantez-Narvaez, J.; Guillen, M.; Alca˜niz, M. Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression. Risks 2019, 7, 70.
6. Abdelhadi, S., ElBahnasy, K.A., Abdelsalam, M.M. A proposed model to predict auto insurance claims using machine learning techniques. . Journal of Theoretical and Applied Information Technology 2020, 98, 3428–3437
7. Jing, L; Zhao, W.; Sharma, K.; Feng, R. Research on Probability-based Learning Application on Car Insurance Data. In 2017 4th International Conference on Machinery, Materials and Computer (MACMC 2017); Amsterdam: Atlantis Press

**WORKLOG**

|  |  |  |
| --- | --- | --- |
| **DAY** | **DATE** | **TASK DONE** |
| **1.** | **27/02/2024** | **Onboarding Process** |
| **2.** | **28/02/2024** | **ID Card Process** |
| **3.** | **05/03/2024** | **Project Setup:**  Gather requirements from stakeholders.  Set up communication channels and tools |
| **4.** | **07/03/2024** | **Data Collection and Preparation:**  Identify relevant datasets  Acquire data from internal databases or external sources |
| **5.** | **12/03/2024** | Perform Data cleaning, including handling missing values, outliers, and inconsistencies.  Explore Data distribution and relationships through visualization and statistical analysis |
| **6.** | **14/03/2024** | **Feature Engineering:**  Extract relevant features from the dataset  Encode categorical variables and handle feature scaling  Generate new features if necessary |
| **7.** | **19/03/2024** | **Model Selection and Training:**  Research Suitable machine learning algorithms for claim prediction  Split the dataset into Training, validation, and test sets |
| **8.** | **21/03/2024** | Evaluate models using appropriate performance metrices |
| **9.** | **26/03/2024** | Turn hyperparameters using techniques like grid search or random search |
| **10.** | **28/03/2024** | **Model Evaluation and Validation:**  Validate models using the test et to assess generalization performances.  Interpret model predictions and provide explanation if necessary |
| **11.** | **02/04/2024** | Analyses Model errors and Biases |
| **12.** | **04/04/2024** | **Deployment and Integration:**  Choose the best performing model for deployment.  Package the model into a deployable format  Integrate the model into existing insurance claim processing system  Implement monitoring and logging for model performance and errors. |
| **13.** | **09/04/2024** | **Documentation and Reporting:** Document the entire process, including data sources, preprocessing steps, model architecture, and deployment procedures.  Create user guides and manuals for stakeholders and end-users.  Prepare a comprehensive report summarizing the project findings, including model performance, insights, and recommendations for future improvements. |
| **14.** | **11/04/2024** | **Maintenance and Iteration:**  Establish a maintenance plan for monitoring model performance and updating as needed.  Collect feedback from end-users and stakeholders for iterative improvements.  Stay updated with new techniques and technologies in machine learning and insurance analytics for continuous enhancement. |

**OFFER LETTER**

****