

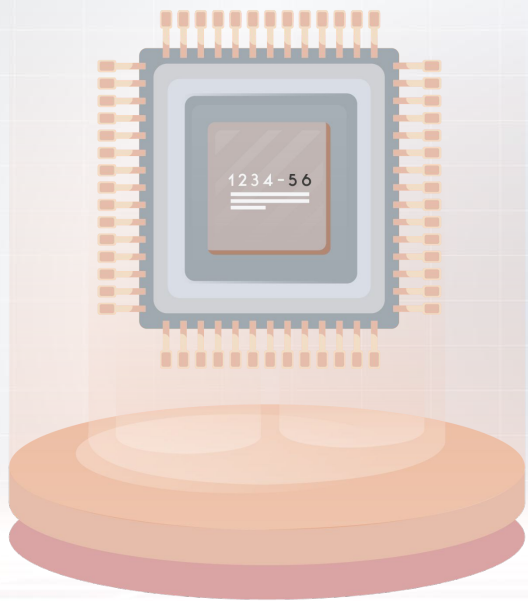


Insurance Claim Prediction Modeling

SC1015 DATA SCIENCE AND AI MINI PROJECT

DIONG WEI CHONG

EE WEN HUI, EVLYN



01

INTRODUCTION

PRACTICAL MOTIVATION

Problems faced by insurance companies:

- ❖ **Risk Management:** Insurance companies face the challenge of effectively managing their risk exposure. They need to accurately identify policyholders who are at a higher risk of filing a claim, allowing them to adjust premiums accordingly or implement proactive measures to mitigate potential risks.
- ❖ **Financial Planning:** One of the key issues for insurance companies is forecasting their financial liabilities accurately. This involves the challenge of predicting claim filings with precision, enabling insurers to establish suitable reserves and efficient allocation of resources to prevent cases of insufficient liquidity

PRACTICAL MOTIVATION



Problems faced by insurance companies:


- ❖ Risk Management
- ❖ Financial Planning

Problem formulation: Classification

- ❖ How can insurance companies increase their prediction accuracy of whether or not a policy holder will file a claim in the next 6 months?

DATA COLLECTION


DATA SOURCE : KAGGLE

 IFTESHA NAJNIN · UPDATED A YEAR AGO

▲ 106

New Notebook

Download (2 MB)



Car Insurance Claim Prediction

Predict whether the policyholder will file a claim in the next 6 months or not.

Data Card

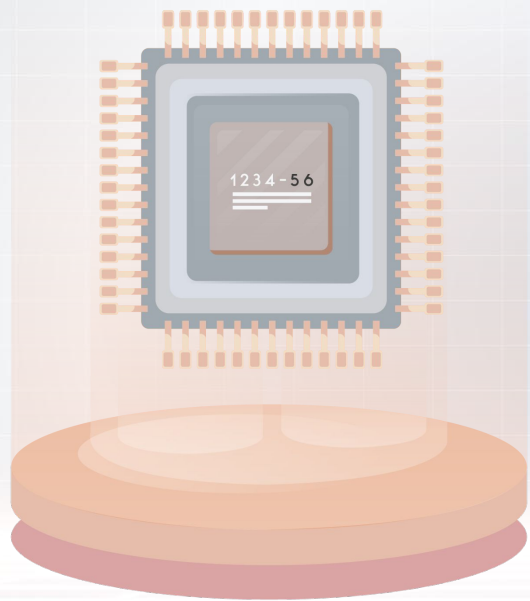
Code (25)

Discussion (4)

Suggestions (0)

Number of insurance coverages: 58 592

Number of variables identified: 44



02

DATA PREPARATION + EDA

DATA PREPARATION

1. Checking for null values in the dataset
 - If there are null values, remove those rows during data cleaning
 - Null values should be removed due to the presence of invalid data
2. Checking for duplicates within the dataset
 - Duplicates should be removed to ensure that the accuracy of the data analyzed is as high as possible
 - If there are duplicates within the dataset, results may be less reliable
3. Understanding the class distribution of the response variable ("is_claim")
 - Finding: Large class imbalance → Needs to be fixed later on to prevent problems for classification models

ANALYSIS OF NUMERIC VARIABLES

AIM: To investigate if there are any numeric variables with a potential correlation with the possibility of an insurance claim

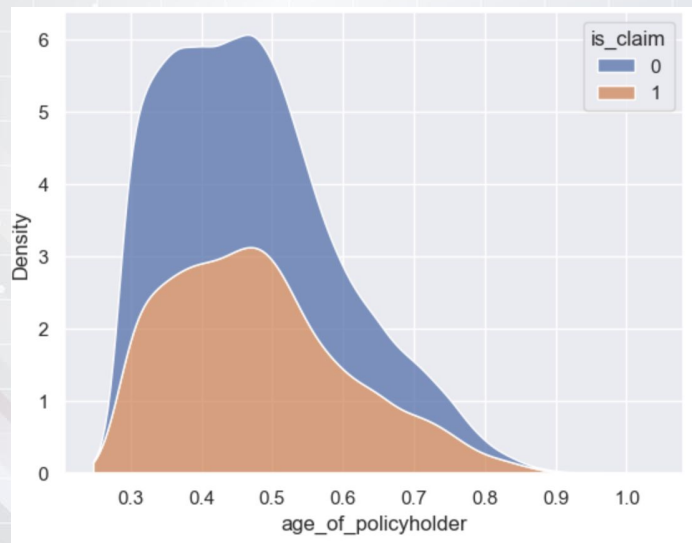
Visualization tools used:

1. Boxplot
 - Useful for summarizing a set of data
 - Shape of boxplot shows the distribution of data, quartiles and outliers
2. Kernel density estimate (KDE) plot
 - Useful method for visualizing the distributions of observations in a dataset
3. Probability histplot
 - Gives the distributions of the probabilities of "is_claim" across different values of a variable

FINDINGS

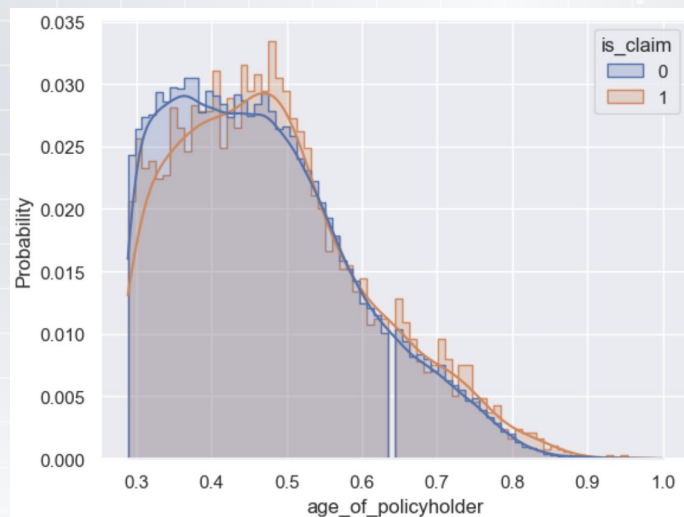
VARIABLE 1: AGE_OF_POLICYHOLDER

1. KDE PLOT



Seems to be a difference in behavior in probability distribution → Possible relationship between “age_of_policyholder” and “is_claim”?

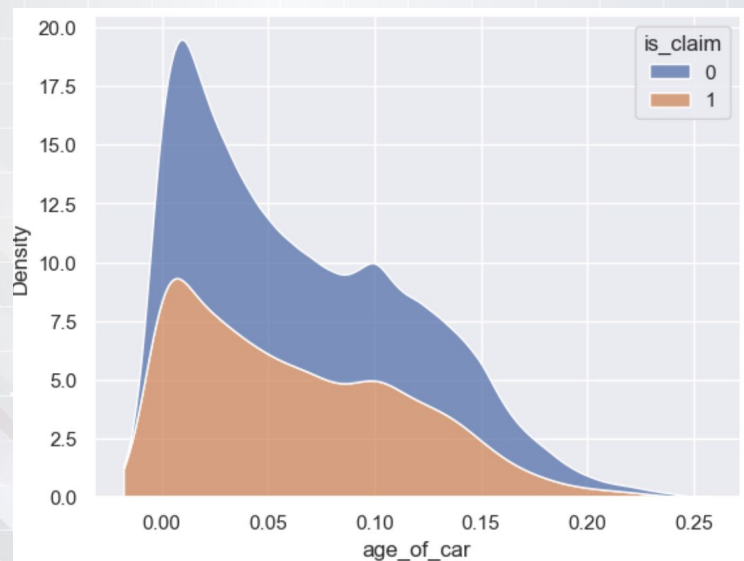
2. HISTPLOT



FINDINGS

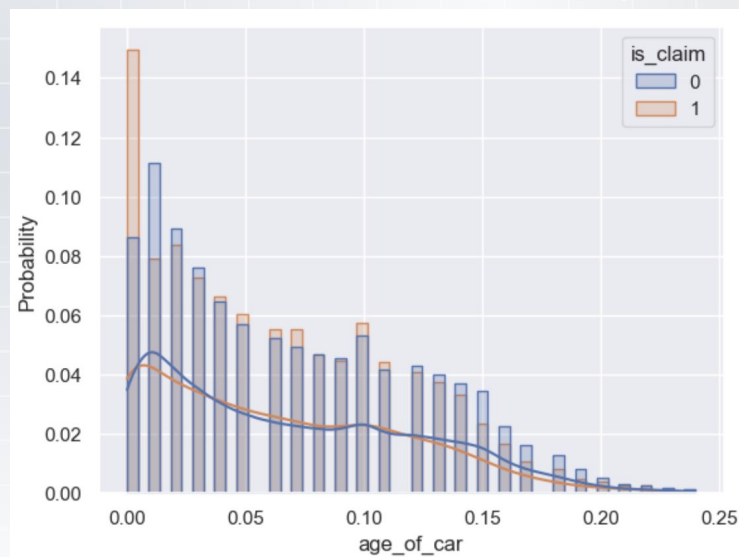
VARIABLE 2: AGE_OF_CAR

1. KDE PLOT



Same distribution between people who have claimed insurance and those who did not claim insurance

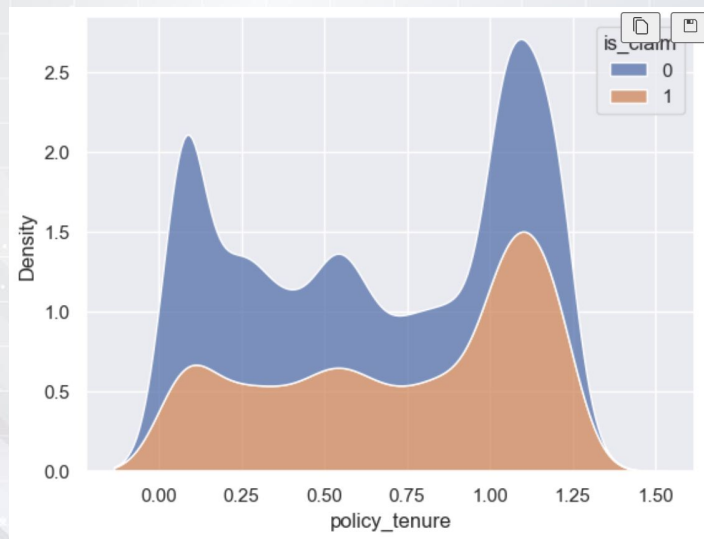
2. HISTPLOT



FINDINGS

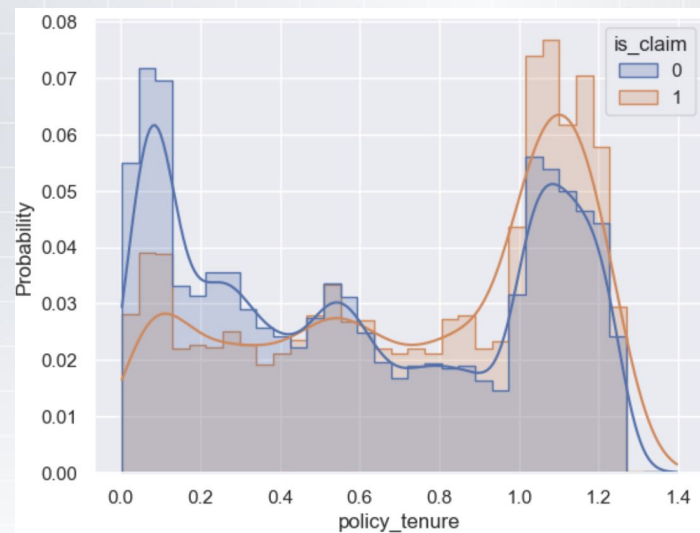
VARIABLE 3: POLICY_TENURE

1. KDE PLOT



Right and left sides behave differently in the histplot → Possible relationship between “policy_tenure” and “is_claim”?

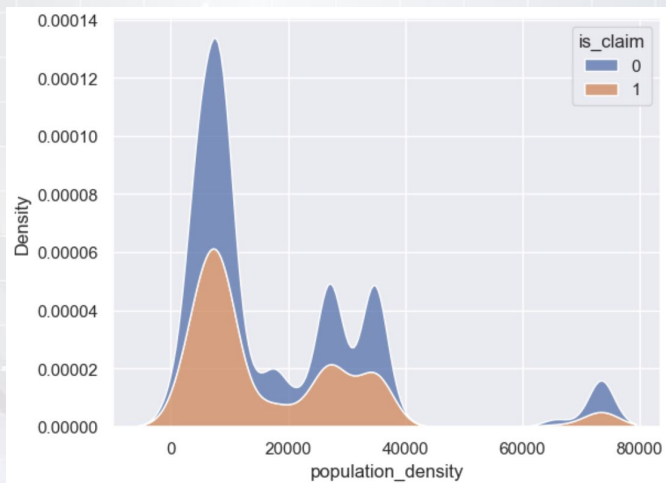
2. HISTPLOT



FINDINGS

VARIABLE 4: POPULATION_DENSITY

1. KDE PLOT



No visible relationship between
"population_density" and "is_claim"

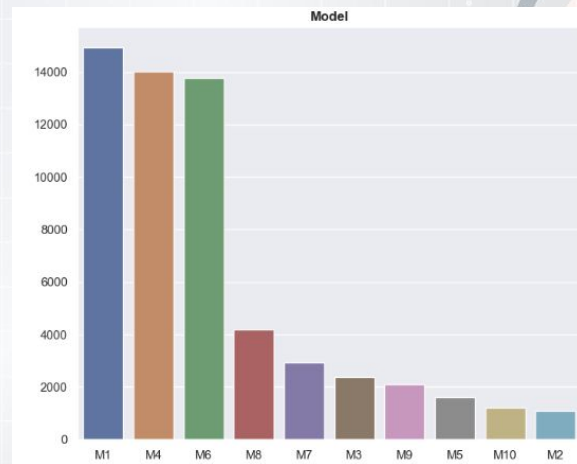
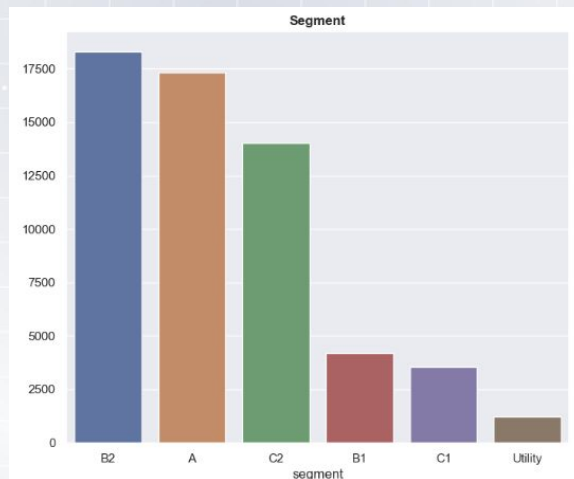
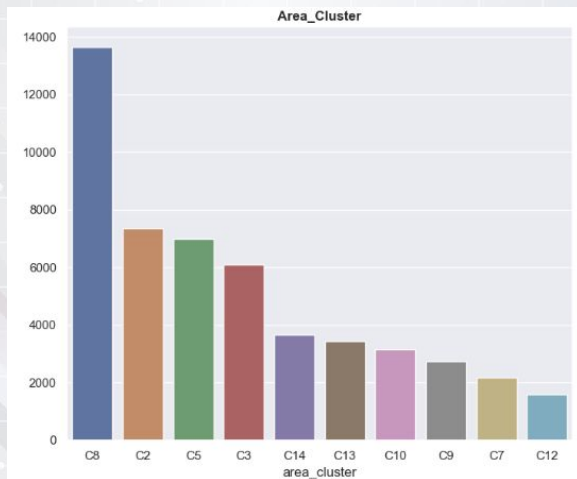
2. LINE GRAPH



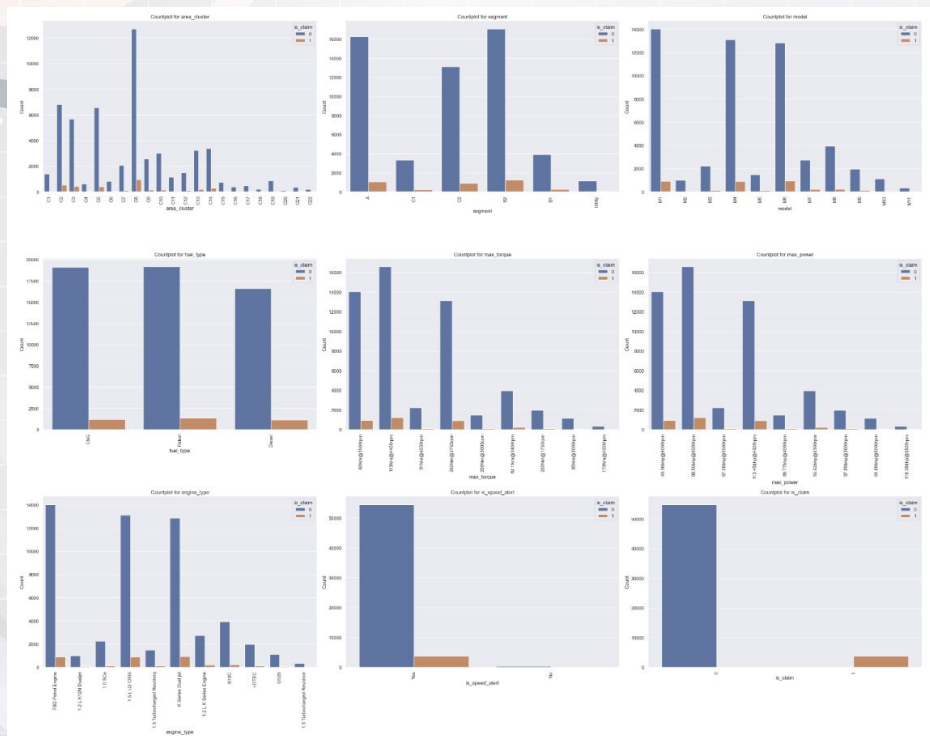
ANALYSIS OF CATEGORICAL VARIABLES

AIM: To investigate if there are any categorical variables with a potential correlation with the possibility of an insurance claim

HISTOGRAM OF THE COUNT FOR EACH CATEGORY



FINDINGS FROM COUNTPLOT



1. Greatest number of claims (~1000) comes from area C8 under "area_cluster"
2. Owners of car models M1, M4 and M6 have the highest number of claims (~1000 each)
3. If the car does not possess a speed alert system, claims = 0

FINDINGS FROM HEATMAP



→ Conducted exploration via heatmaps to determine the correlation between different variables and the likelihood of a claim

→ Corroborates some of our findings from the countplot (eg. Identification of M1, M4 and M6 car models)

OVERALL FINDINGS

TOP 5 VARIABLES TO EXPLORE:

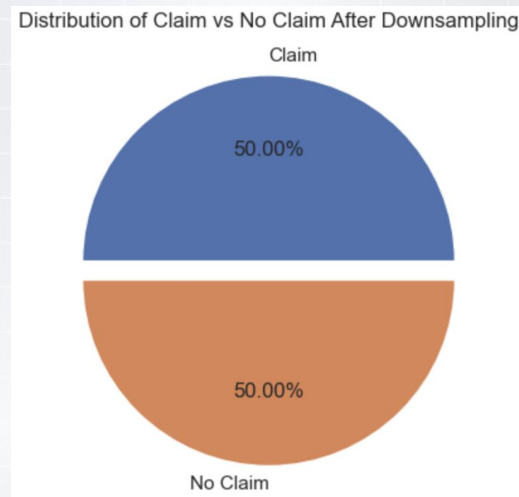
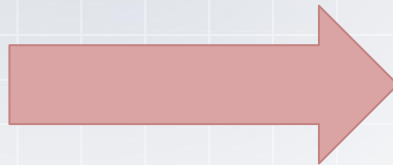
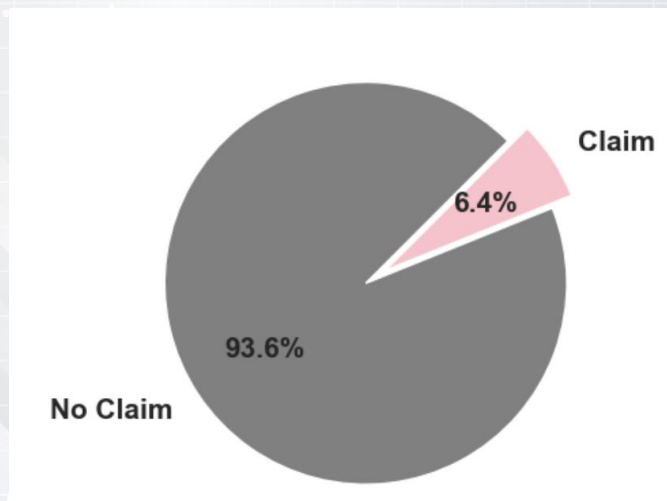
1. AGE OF POLICYHOLDER
2. AGE OF CAR
3. DURATION OF POLICY TENURE
4. PRESENCE OF SPEED ALERT
5. MODELS OF CAR

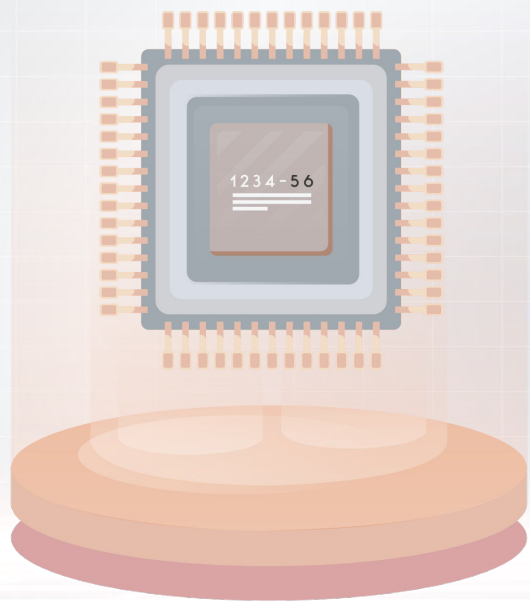


DOWNSAMPLING

→ LARGE CLASS IMBALANCE IN OUR RESPONSE VARIABLE "IS_CLAIM", WE WILL NEED TO CONDUCT DOWNSAMPLING.

Initial ratio of no claims to claims: 14.63





03

MACHINE LEARNING

Overall ML Approach

Binary Classification

Using **both numerical and categorical features** identified as relevant in the EDA about the policy holder and their car such as the age of the insurance policy holder, weight of car, length of car etc to classify if the policy holder will file a claim in the next 6 months

Categorical Features were pre-processed using One Hot Encoder for analysis

Subsequently, the data was split into train set and test set to begin model building

Models Explored



1. **Random Forest Regression Model**
 - a. Uses an ensemble of decision trees, which are each built using a random subset of overall features, which contributes to diversity among trees
2. **Logistic Regression Model**
 - a. Estimates the parameters of the model (by attributing a weight to the included features)
3. **Naive Bayes Model**
 - a. Features are considered independently using Bayes Theorem. Unlike discriminative classifiers such as logistic regression, it doesn't learn which features are most crucial for distinguishing between classes
4. **Voting Classifier (Hard and Soft)**
 - a. Uses an ensemble of the above three models to increase predictive accuracy



Random Forest Regression

1. Hyper parameters of the number of trees, max depth of tree and number of features to be included in each decision tree can greatly affect test accuracy
2. Initial model used 50 trees with a max depth of 3. Test accuracy of **0.59** was achieved

TPR Train : 0.7455516014234875

TNR Train : 0.5455192034139402

FPR Train : 0.4544807965860597

FNR Train : 0.25444839857651247

Test Accuracy : 0.5949839914621131

How do we decide **suitable hyper parameters** for a Random Forest Regression?

GridSearch CV

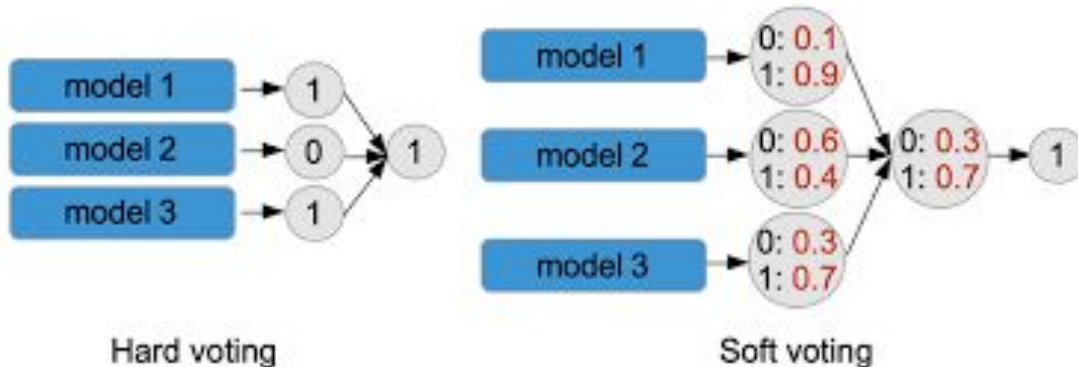
1. Implemented a Grid Search of different combinations of hyperparameters to improve classification accuracy
2. Model would iterate through different number of trees, different max depths and different number of features included to find the best model with the highest test accuracy score

```
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import accuracy_score
#Define the hyperparameter grid
p_grid = {
    'n_estimators': [50, 100, 200, 300],
    'max_depth': [3, 4, 5,],
    'max_features': ['sqrt', 'log2']
}
```


Voting Classifier

1. Uses the ensemble of the three models to predict the class label (claim or no claim) by taking a vote

```
voting_clf_hard = VotingClassifier(  
    estimators=[  
        (labels[0], clf1), # Include the first classifier (Logistic Regression)  
        (labels[1], clf2), # Include the second classifier (Random Forest)  
        (labels[2], clf3), # Include the third classifier (Naive Bayes)  
    ],  
    voting='hard' # Specify hard voting, where the majority class prediction is chosen  
)
```



Overall Accuracy of Models

1. Accuracy of all models remain low, with Random Forest Regression having the highest test accuracy
2. Surprisingly, the Voting Classifier has a lower accuracy. This may be because all three models may be too highly correlated and hence there is a lack of diversity in the base model used. Furthermore, initial low accuracy of models used may further inhibit the voting classifier
3. Possible reason for low accuracy is that there is too much 'noise' or irrelevant features in the data set that reduce the model's accuracy

Accuracy: 0.58 [Logistic Regression]

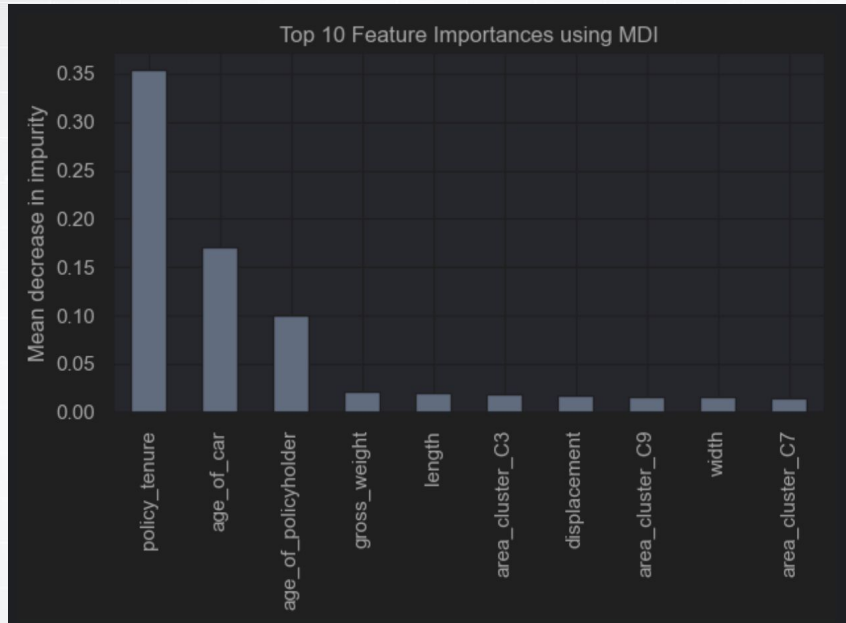
Accuracy: 0.61 [Random Forest]

Accuracy: 0.53 [Naive Bayes]

Accuracy: 0.58 [Voting_Classifier_Hard]

Accuracy: 0.54 [Voting_Classifier_Soft]

Feature Importance and Noise reduction

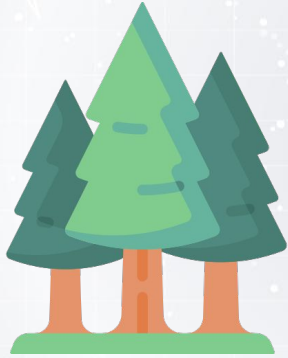


Test Data

Accuracy : 1.0

1. Using the Random Forest Model, we plotted the **top ten features that reduced the gini impurity**
2. As expected from our EDA, the top three features were **AGE OF POLICYHOLDER, AGE OF CAR** and **DURATION OF POLICY TENURE**
3. To reduce noise, we built another random forest model using the top 5 features only
4. Testing accuracy greatly increased to 1.0

Outcome of Project



Successfully identified
Random Forest Model as
our preferred model due to
its high level of predictive
accuracy



Narrowed down to **5**
required features for
insurance companies to
classify whether policy
holders will file a claim soon
or not

Model



Selected suitable **model hyperparameters** using GridSearch CV.

However, there is a tradeoff as models with more trees or greater depth of trees increases the time complexity and space complexity of the model

Deeper analysis into the model such as if it overfits the data or presence of high cardinality bias where model attributes more importance to features with many unique values can be done to further fine tune the model

Recommendations



1. Choose **not to provide coverage or charge higher premiums** to new policy applicants that are likely to file a claim in the next 6 months as company will likely suffer a loss on this group of people
2. Collect **only selected data points required for the model from policyholders** to reduce information costs
3. Allow companies to **predict amount of potential future insurance claims** so as to better manage employee workload and work plan forecasting