Prediction of Potential Vehicle Insurance Fraud

Team 05

Ying Chai
Jingping Guo
Jing Lin
Yifei Li
Shenger Zhang

AGENDA

- Introduction
- Research Question
- Data Processing
- Modeling
- Analysis and Finding
- Model Selection
- Conclusion

INTRODUCTION

ABOUT VEHICLE INSURANCE FRAUD

$1,000
average household pays

$80 Billion
Stolen in US annually

$29 Billion
Insurance companies lost

1893 Cases
in Maryland annually

What are we interested in?

What the target features are ?

How to improve the insurance screening rate?

What algorithms do we need to use?

Feature Data

- 33 features including target variable
- Sample size: 15,420
    - Training 70%, Testing 30%
- Most of them are categorical
- Fraud are classified into 2 classes
    - Fraudulent and Not Fraudulent

```
Rows: 15,420
Columns: 33
$ Month                <chr> "Dec", "Jan", "Oct", "Jun", "Jan", "Oct", "Feb", "Nov", "Dec", "Apr", "Mar", "Mar", '
$ WeekOfMonth          <int> 5, 3, 5, 2, 5, 4, 1, 1, 4, 3, 2, 5, 3, 5, 5, 4, 4, 5, 4, 4, 2, 2, 3, 3, 3, 3, 3, 3, 1
$ DayOfWeek            <chr> "Wednesday", "Wednesday", "Friday", "Saturday", "Monday", "Friday", "Saturday", "Fric
$ Make                 <chr> "Honda", "Honda", "Honda", "Toyota", "Honda", "Honda", "Honda", "Honda", "Honda", "Fo
$ AccidentArea         <chr> "Urban", "Urban", "Urban", "Rural", "Urban", "Urban", "Urban", "Urban", "Urban", "Urb
$ DayOfWeekClaimed     <chr> "Tuesday", "Monday", "Thursday", "Friday", "Tuesday", "Wednesday", "Monday", "Tuesday
$ MonthClaimed         <chr> "Jan", "Jan", "Nov", "Jul", "Feb", "Nov", "Feb", "Mar", "Dec", "Apr", "Mar", "Mar", '
$ WeekOfMonthClaimed   <int> 1, 4, 2, 1, 2, 1, 3, 4, 5, 3, 3, 5, 3, 1, 1, 5, 1, 1, 5, 1, 1, 2, 5, 3, 3, 1, 4, 4, 4
$ Sex                  <chr> "Female", "Male", "Male", "Male", "Female", "Male", "Male", "Male", "Male", "Male", '
$ MaritalStatus        <chr> "Single", "Single", "Married", "Married", "Single", "Single", "Married", "Single", "S
$ Age                  <int> 21, 34, 47, 65, 27, 20, 36, 0, 30, 42, 71, 52, 28, 0, 61, 38, 41, 28, 32, 30, 40, 47,
$ Fault                <chr> "Policy Holder", "Policy Holder", "Policy Holder", "Third Party", "Third Party", "Thi
$ PolicyType           <chr> "Sport - Liability", "Sport - Collision", "Sport - Collision", "Sedan - Liability", '
$ VehicleCategory      <chr> "Sport", "Sport", "Sport", "Sport", "Sport", "Sport", "Sport", "Sport", "Sport", "Uti
$ VehiclePrice         <chr> "more than 69000", "more than 69000", "more than 69000", "20000 to 29000", "more thar
$ PolicyNumber         <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24
$ RepNumber            <int> 12, 15, 7, 4, 3, 12, 14, 1, 7, 7, 7, 13, 11, 12, 3, 16, 15, 6, 6, 2, 3, 13, 8, 5, 12,
$ Deductible           <int> 300, 400, 400, 400, 400, 400, 400, 400, 400, 400, 400, 400, 400, 400, 400, 400, 400,
$ DriverRating         <int> 1, 4, 3, 2, 1, 3, 1, 4, 4, 1, 3, 1, 1, 3, 1, 1, 4, 1, 1, 2, 1, 2, 3, 3, 3, 4, 2, 3, 1
$ Days_Policy_Accident <chr> "more than 30", "more than 30", "more than 30", "more than 30", "more than 30", "more
$ Days_Policy_Claim    <chr> "more than 30", "more than 30", "more than 30", "more than 30", "more than 30", "more
$ PastNumberOfClaims   <chr> "none", "none", "1", "1", "none", "none", "1", "1", "none", "2 to 4", "none", "2 to 4
$ AgeOfVehicle         <chr> "3 years", "6 years", "7 years", "more than 7", "5 years", "5 years", "7 years", "new
$ AgeOfPolicyHolder    <chr> "26 to 30", "31 to 35", "41 to 50", "51 to 65", "31 to 35", "21 to 25", "36 to 40", '
$ PoliceReportFiled    <chr> "No", "Yes", "No", "Yes", "No", "No", "No", "No", "No", "No", "No", "No", "No", "No",
$ WitnessPresent       <chr> "No", "No", "No", "No", "No", "No", "No", "No", "Yes", "No", "No", "No", "No", "No",
$ AgentType            <chr> "External", "External", "External", "External", "External", "External", "External", '
$ NumberOfSuppliments  <chr> "none", "none", "none", "more than 5", "none", "3 to 5", "1 to 2", "none", "3 to 5",
$ AddressChange_Claim  <chr> "1 year", "no change", "no change", "no change", "no change", "no change", "no change
$ NumberOfCars         <chr> "3 to 4", "1 vehicle", "1 vehicle", "1 vehicle", "1 vehicle", "1 vehicle", "1 vehicle
$ Year                 <int> 1994, 1994, 1994, 1994, 1994, 1994, 1994, 1994, 1994, 1994, 1994, 1994, 1994, 1994, 1
$ BasePolicy           <chr> "Liability", "Collision", "Collision", "Liability", "Collision", "Collision", "Collis
$ FraudFound_P         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1
```

Data Source

- Kaggle: Vehicle Insurance Fraud Detection

  https://www.kaggle.com/datasets/shivamb/vehicle-claim-fraud-detection

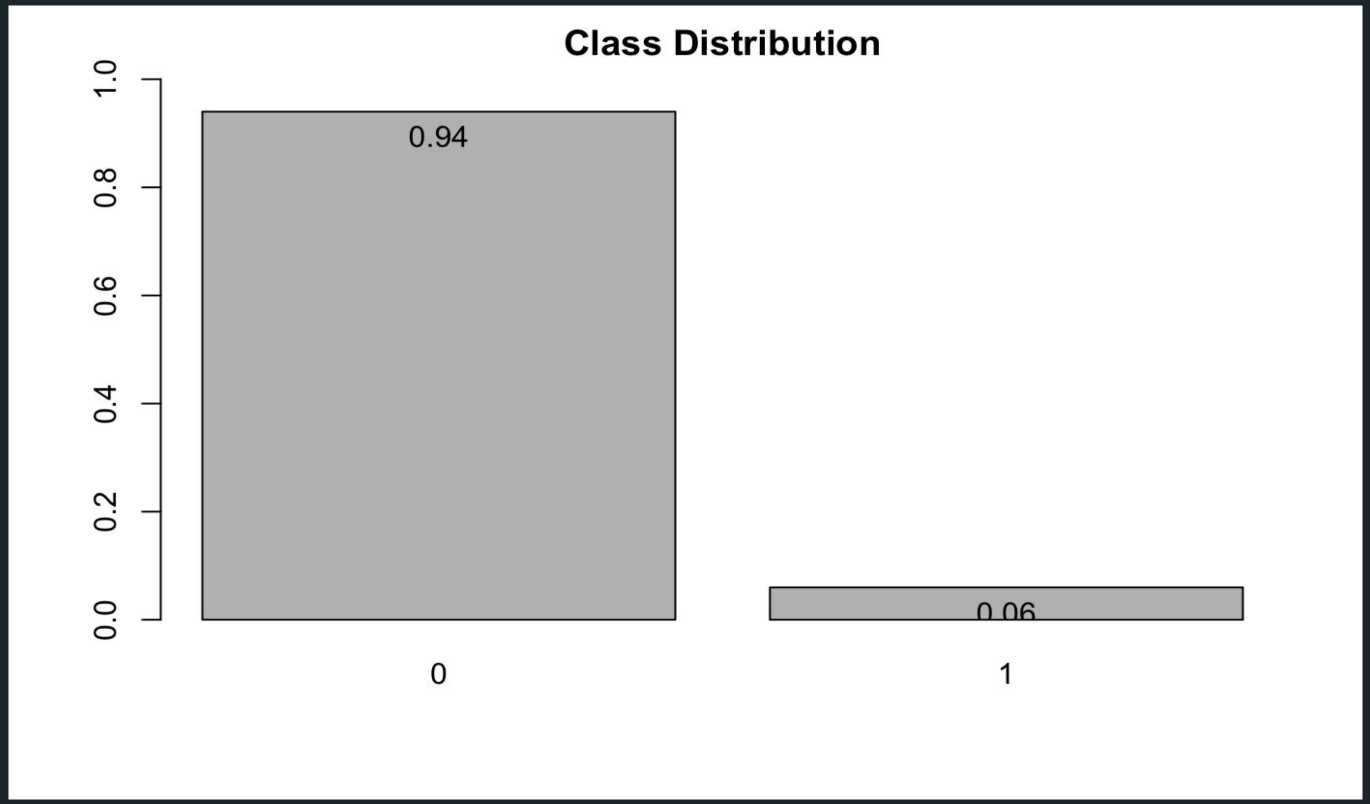
    - Original data: Oracle

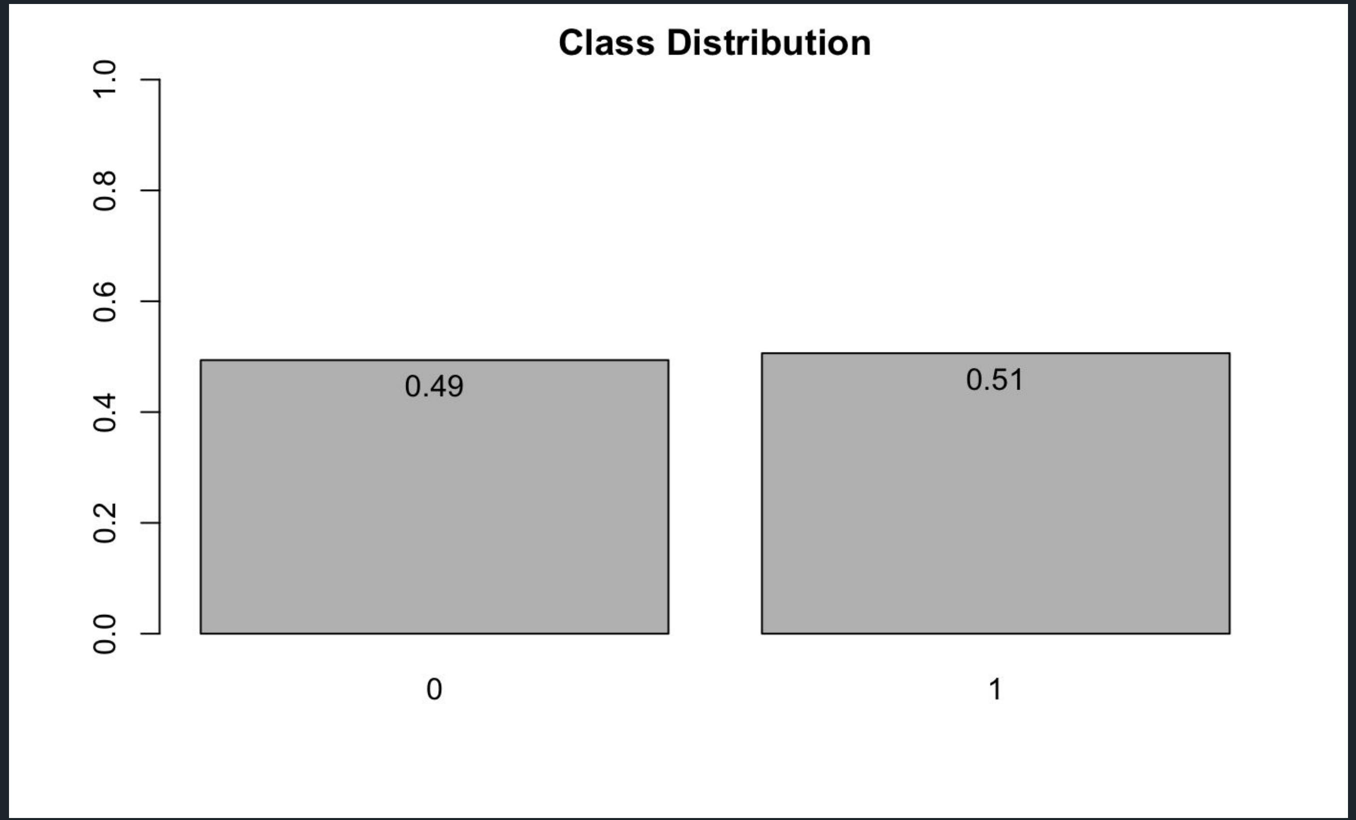
Data Pre-process

- Remove records with any missing values and Replace invalid data
    - 1 Record DayOfWeekClaimed = 0 & MonthClaimed = 0
    - 320 Records (Age = 0)
- Convert all character type to factor

Imbalance Data

- Class Proportion: 6% of Fraud and Not Fraud 94%
- Upsample training data
    - Duplicating the minority class many times





Methodology

WHAT MODEL DID WE USE

Logistic Regression	Classification Tree	Random Forest	Boosting	XGBoost
MODEL 1	MODEL 2	MODEL 3	MODEL 4	MODEL 5

Model1

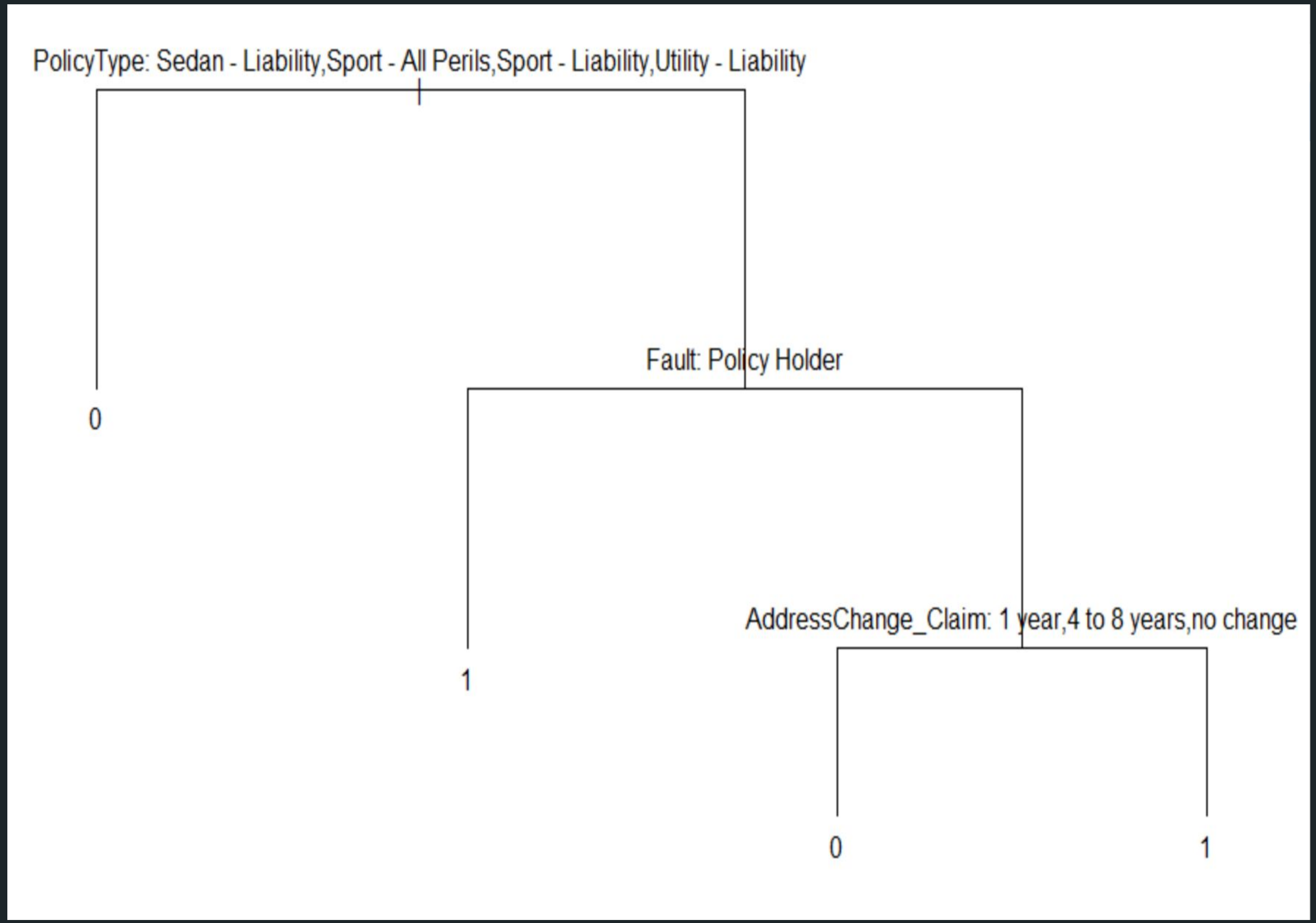
LOGISTIC REGRESSION

- Upsampling works
- Low accuracy (0.65)
- High sensitivity (0.83)

Analysis and Finding(Cont.)

Model 2

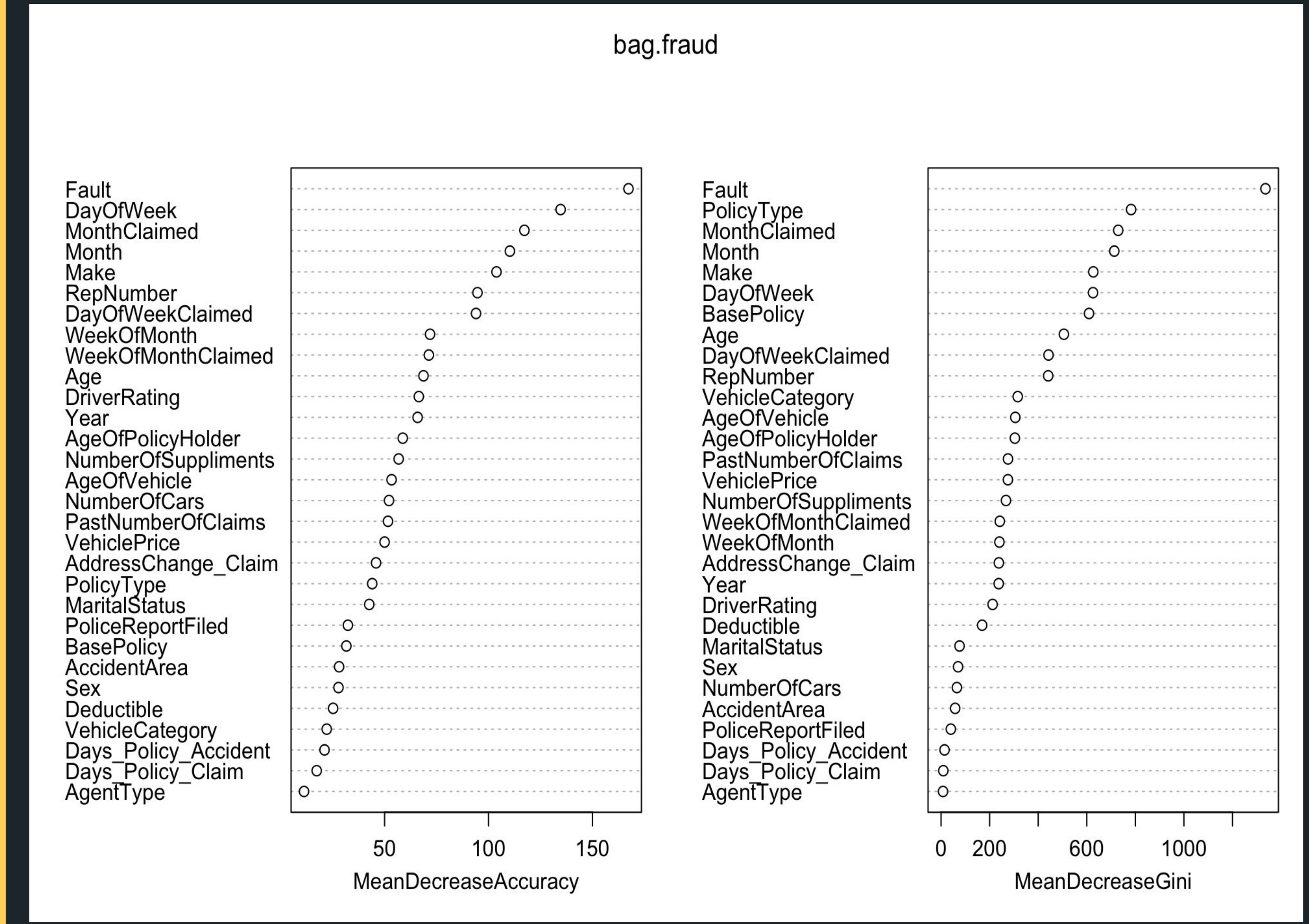
CLASSIFICATION TREE



- Growing a full tree -> Pruned tree
- Low Accuracy (0.6)
- High sensitivity (0.96)

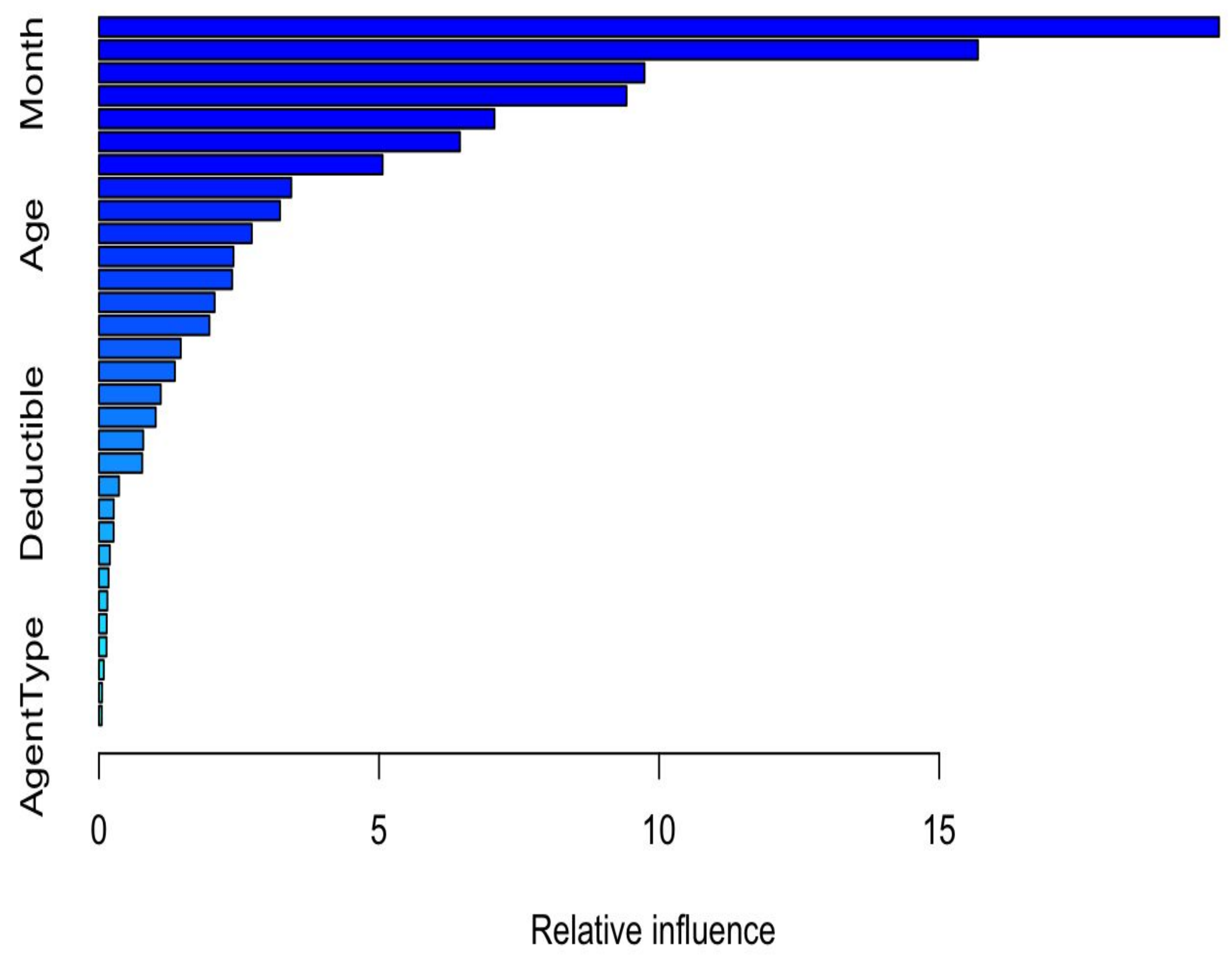
Model3

RANDOM FOREST



- Intelligent algorithm: comprehensive & randomness
- High Accuracy (0.94)
- High sensitivity (0.05): Not applicable in reality

Analysis and Finding(Cont.)



Model 4

BOOSTING

- "upweights" misclassified data points
- parameter: learning rate=0.001(default)

  tree number=5000
- High accuracy(0.93), low sensitivity(0.21)
- important variables

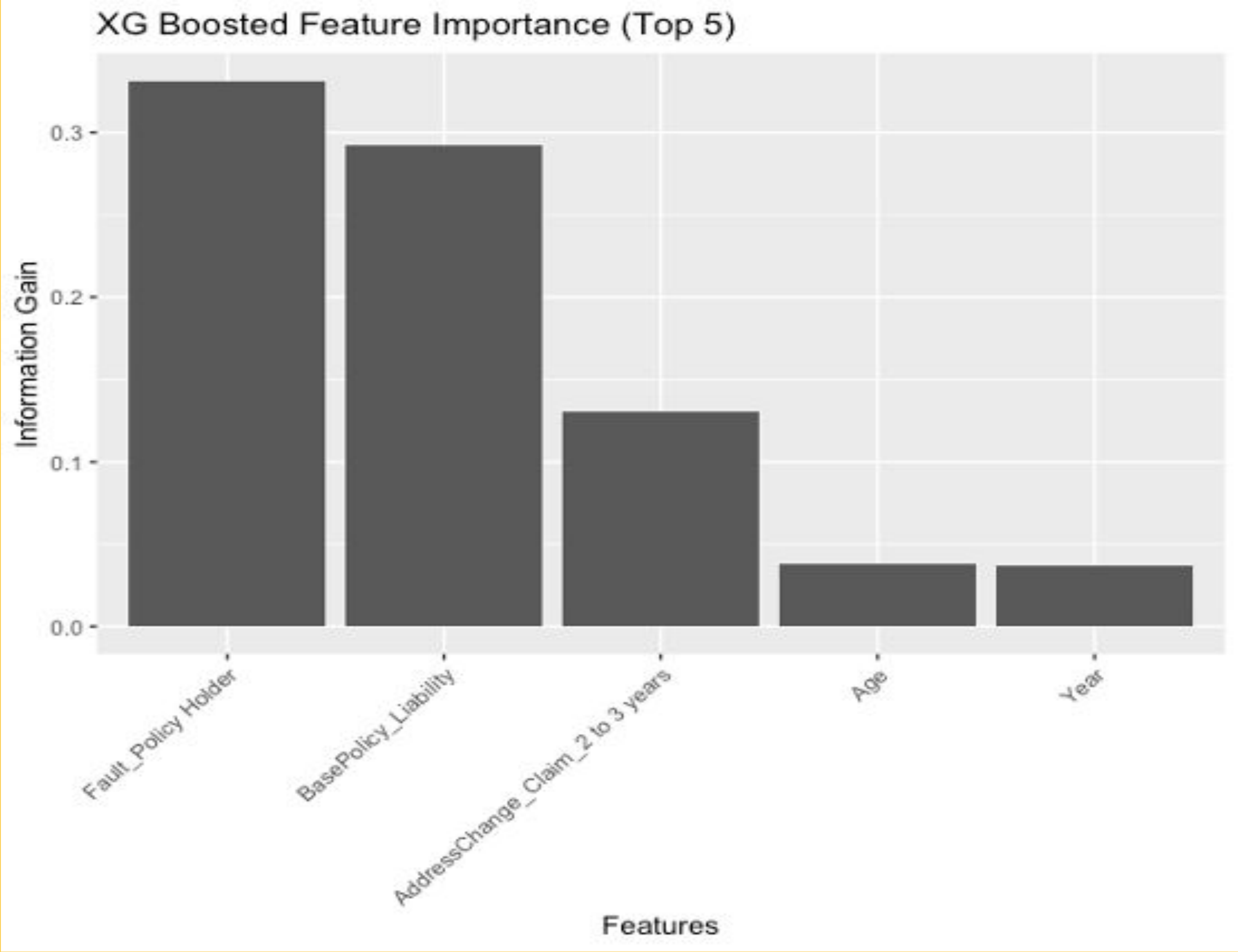
Model 5

XGBOOST

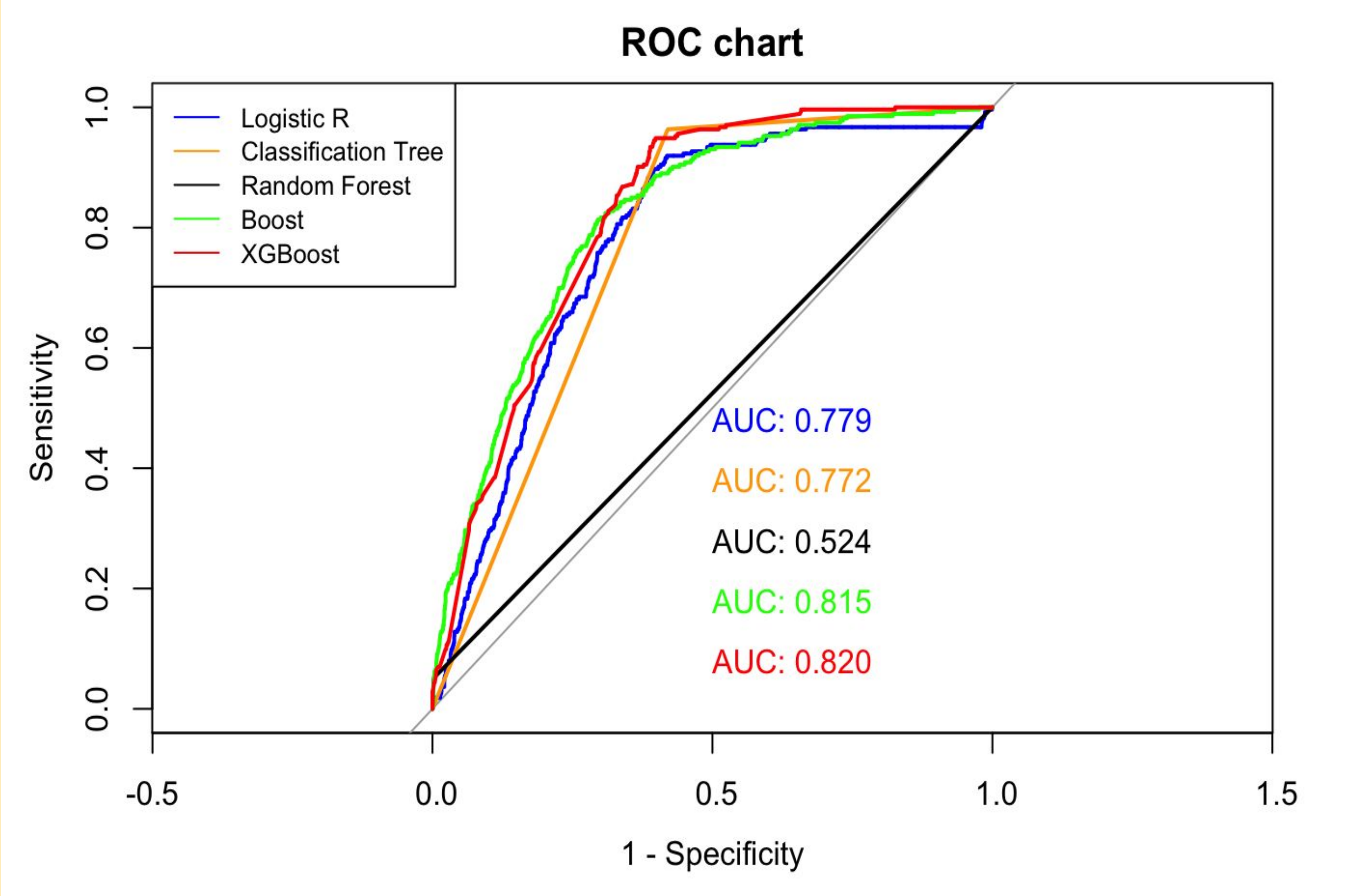
- regularization techniques
- parameter:

  logistic regression for classification

  maximum depth of the tree=4
- Low accuracy(0.66), high sensitivity(0.87)
- important variables



Model	Upsample Ratio: 1 : 1 (class 0 vs class 1)			
	Accuracy	Specificity	Sensitivity	AUC
Logistic	0.6484	0.6369	0.8315	0.779
Class Full tree	0.6024	0.5797	0.9634	
Pruned tree	0.6024	0.5797	0.9634	0.772
Random Forest	0.9418	0.9977	0.0512	0.524
Boosting	0.9258	0.9710	0.2051	0.815
XGboost	0.6595	0.6461	0.8718	0.820



Sensitivity best: Classification Tree
Accuracy best: Random Forest
Overall best (Relatively) : XGBoost

Model Selection

Conclusions

1 **We selected the best predictive models based on different indicators:**

Sensitivity best: Classification Tree
Accuracy best: Random Forest
Overall best : XGBoost

We figured out some important features that probably affect a claim is fraud or not: 2

Fault
Which month it’s claimed
The make of car
Which day of week
The policy type

Improvement

- To find a better way when upsampling instead of simply duplicating data
- To find a method to merge some features or decrease some categories (too many features lead to reduced validity）
- To prove whether or not there is a better way to trade off the accuracy and sensitivity in this case (more model to try)

THANKS

FOR YOUR

LISTENING!