یک الگوریتم یادگیری ماشین با نظارت برای شناسایی و پیش‌بینی تقلب در تراکنش‌های کارت اعتباری

A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions

Stu. : Mahdi Mahdiani
Prof. : Dr Rezvanian
Machine Learning Presentation-Dey 1402

سرفصل مطالب

مقدمه

- ارائه خدمات شخصی و حضوری تا سال ۱۹۹۶
- معرفی بانکداری اینترنتی توسط Citibank و Fargo Bank
- تغییر روش مدیریت پول در زندگی روزانه
- امکان تراکنش مالی به صورت اینترنتی از خانه یا دفتر کار
- چالش کاهبرداری کارت های اعتباری در تراکنش های آنلاین
- رشد قابل توجه میزان جرایم مربوط به تراکنش های مالی
- ضرر ۱.۴۶ میلیون دلاری در بانک غنا

پیشینه پژوهش

ردیف	سال	الگوریتم/مدل/روش	توضیحات	نتیجه
۱	۲۰۲۲	XGboost	استفاده از روش XGboost	**AUC : 99**
۲	۲۰۱۸	AdaBoost Majority Voting	در این مقاله از این دو الگوریتم استفاده شد با اضافه کردن نویز ۱۰ ، ۲۰ ، ۳۰ درصد	الگوریتم Majority Voting به عنوان الگوریتم برتر معرفی شد
۳	۲۰۱۹	LR,NB,RF,Multilayer perceptron	استفاده از روش SMOTE و الگوریتم های ذکر شده	LR :97.46 NB:99.23 RF:99.96 MLP:99.93
۴	۲۰۰۸	Hidden Markov Model	استفاده از Baum-welch برای تعیین پارامتر های مدل مارکوف	Accuracy : 80
۵	۲۰۲۱	Deep Convolution Neural Network	memory based deep learning neural network	Accuracy : 99.96 بعد از ۱۰۰۰

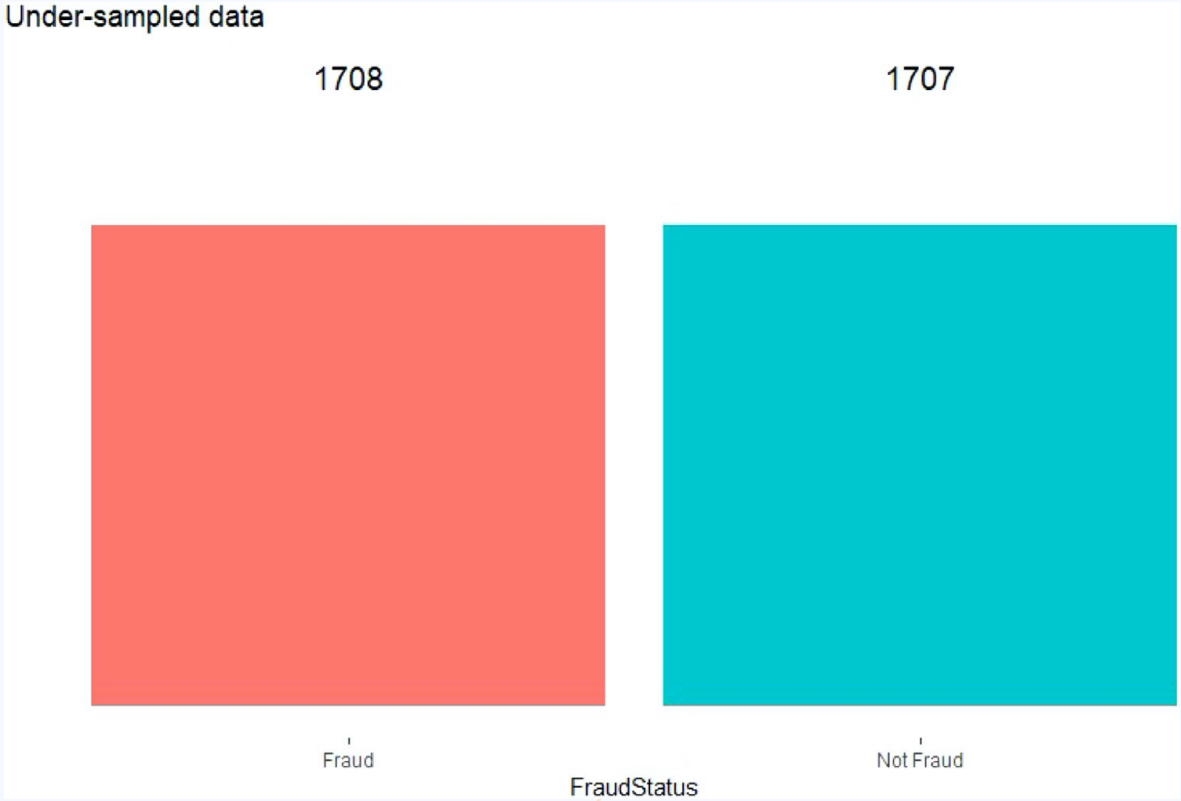
داده ها

- تراکنش های شبیه سازی شده سال ۲۰۲۰ شامل تراکنش های تقلبی(کلاهبرداری) و عادی
- شامل اطلاعاتی مثل نام مشتری , فروشنده , نوع خرید , طبقه بندی و ...
- دیتاست دارای ۵۵۵۷۱۹ ردیف و ۲۳ ستون است
- ۱۲ مورد از متغیر ها کیفی است

پیش پردازش داده ها

1. تمیز کردن داده ها و حذف داده های گم شده
2. با استفاده از feature scaling داده های عددی را بین ۰ تا ۱ قرار دادن
3. نمونه گیری از داده های نامتعادل جهت جلوگیری از سوگیری به سمت اکثریت
4. استفاده از تکنیک SMOTE

نمونه گیری



خلاصه آماری

Table 1
Basic Statistics for the character variables.

Name	Count	Unique	Top	Frequency
Transaction date and time	555 719	544 760	2020-12-19 16:02:22	4
Merchant	555 719	693	fraud_Kilback LLC	1859
Category	555 719	14	gas_transport	56 370
First	555 719	341	Christopher	11 443
Last	555 719	471	Smith	12 146
Gender	555 719	2	F	304 886
Street	555 719	924	444 Robert Mews	1474
City	555 719	849	Birmingham	2423
State	555 719	50	TX	40 393
Job	555 719	478	Film/video editor	4119
Date of birth	555 719	910	1977-03-23	2408
Transaction number	555 719	555 719	2da90c7d74bd46a	1

خلاصه آماری

Table 2
Basic Statistics for the numeric variables.

Name	Count	Mean	Std	Min	25%	50%	75%	Max
Unique identifier	555 719	277 859	160 422.4	0	138 929.5	277 859	416 788.5	555 718
Credit card number of customers	555 719	4 178 387	1 309 837	6 041 621	1 800 429	3 521 417	4 635 331	4 992 346
Amount	555 719	69.39	156.75	1	9.63	47.29	83.01	22 768.11
Zip	555 719	48 842.63	26 855.28	1257	26 292	48 174	72 011	99 921
Latitude	555 719	38.54	5.061	20.03	34.67	39.37	41.89	65.69
Longitude	555 719	−90.23	13.72	−165.67	−96.8	−87.48	−80.18	−67.95
City population	555 719	88 221.89	300 390.9	23	741	2408	19 685	2 906 700
Time (s)	555 719	1 380 679	5 201 104	1 371 817	1 376 029	1 380 762	1 385 867	1 388 534
Merchant latitude	555 719	38.54	5.1	19.03	34.76	39.38	41.95	66.68
Merchant longitude	555 719	−90.23	13.73	−166.67	−96.91	−87.45	−80.27	−66.95
Fraud status	555 719	0.0039	0.062	0	0	0	0	1

روش ها

1. Decision Tree
2. Logistic Regression
3. Random Forest

درخت تصمیم

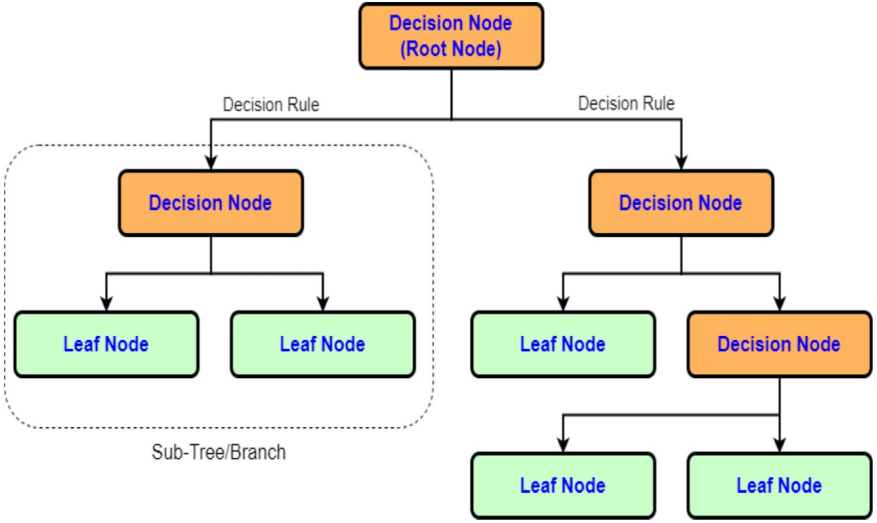


Fig. 2. Decision tree.

- یک الگوریتم یادگیری با نظارت برای طبقه بندی است
- یک ساختار درختی برای تصمیم گیری تولید می کند
- اجزای اصلی آن گره ها و لبه ها هستند
- گره های برگ به عنوان نمایشی از کلاس ها برای طبقه بندی عمل می کنند

درخت تصمیم(ادامه)

Entropy

$$E(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \quad (1)$$

$$E(X) = -p(Fraud) \log_2 p(Fraud) - p(Not\ Fraud) \log_2 p(Not\ Fraud) \quad (2)$$

Gini

$$E(X) = 1 - \sum_{i=1}^{n} p(x_i)^2$$

Information Gain

$$G(X,Y) = E(X) - E(X|Y) \quad (4)$$

$$G(X,Y) = -p(Fraud) \log_2 p(Fraud) - p(Not\ Fraud) \log_2 p(Not\ Fraud) - \sum \frac{|Sv|}{S} entropy(Sv) \quad (5)$$

رگرسیون لجستیک

Formula

$$\ln\left[\frac{p(y=1)}{1-p(y=1)}\right] = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$$

where;

α_0 is the intercept of the model

α_i are the model coefficients , $i = 1, 2, 3, \ldots, n$

x_i are the independent variables, $i = 1, 2, 3, \ldots, n$

y is the dichotomous dependent variable

$$p(y) = \begin{cases} 1, & fraud \\ 0, & no\ fraud \end{cases}$$

Using Sigmoid Function

$$p(y) = \frac{exp(\alpha + \beta x)}{1 + exp(\alpha + \beta x)}$$

منحنی یادگیری رگرسیون لجستیک

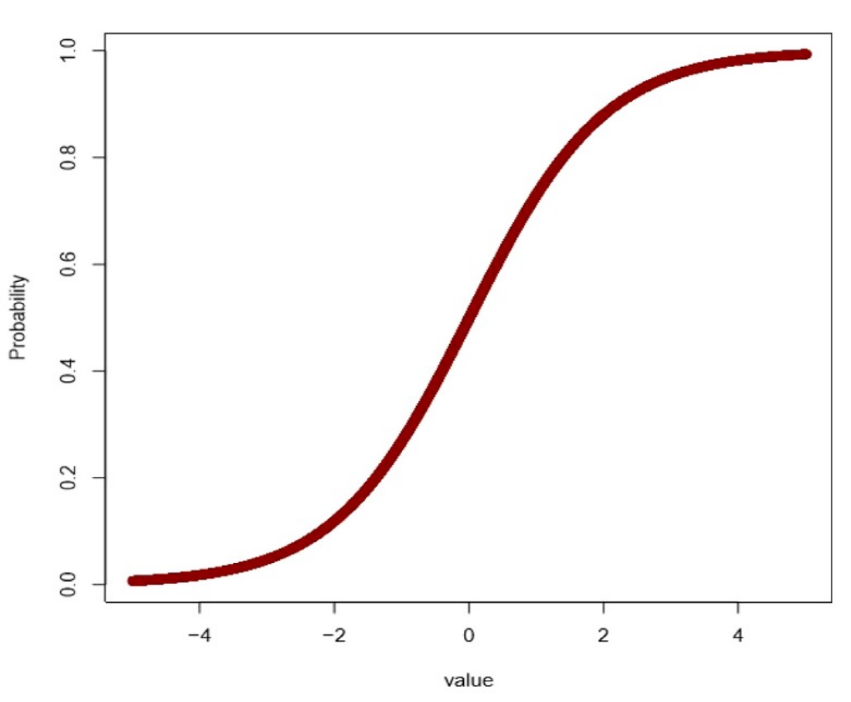


Fig.3 Learning Curve

جنگل تصادفی

۱. تعریف و هدف
- الگوریتم یادگیری با نظارت
- استفاده از گروهی از مدل های درخت تصمیم برای طبقه بندی

۲. Weak Learners

۳. Ensemble Learning

۴. Bagging Method

۵. Forest of Decision tree

$$D(X) = \arg\max \left\{ \sum_{i}^{N} dK_i(X) = Fraud, \sum_{i}^{N} dK_i(X) = Not\ fraud, \right\}$$

جنگل تصادفی(ادامه)

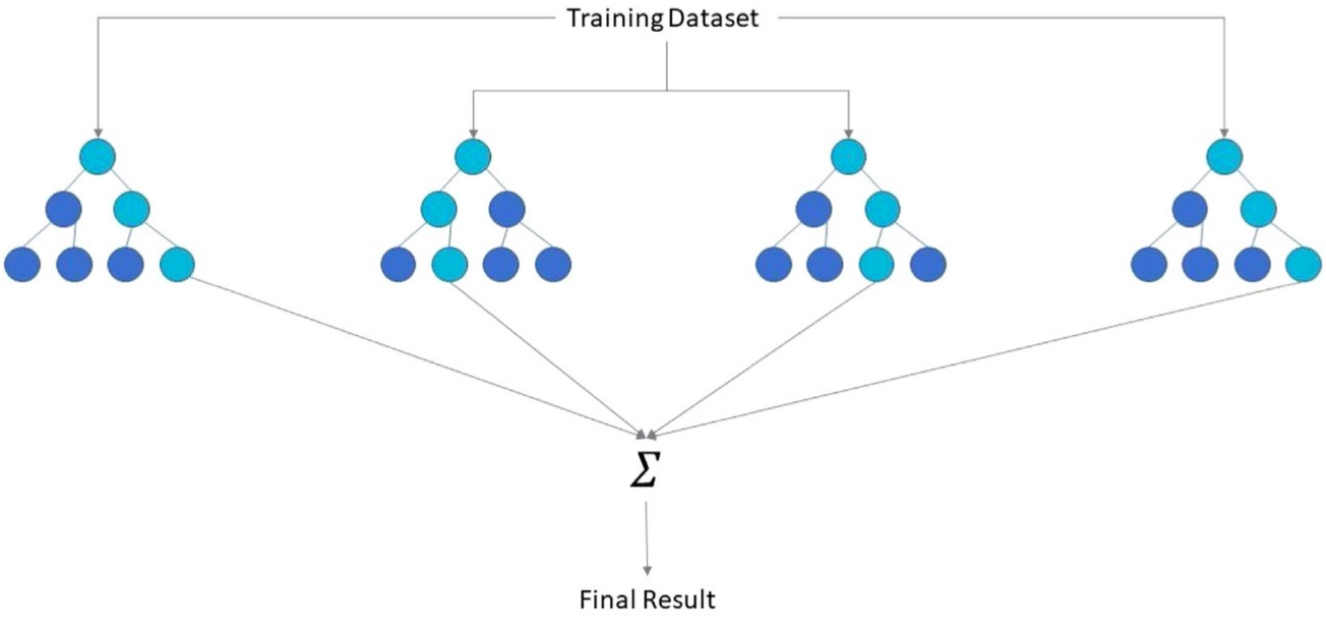


Fig. 4 Random forest.

معیارهای عملکرد

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP}$$

$$Precision = \frac{TP}{FP + TP}$$

$$Recall/Sensitivity = \frac{TP}{TP + FN}$$

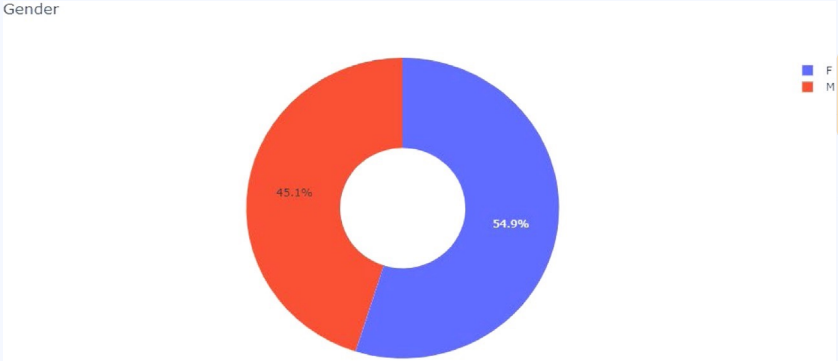
$$Specificity = \frac{TN}{TN + FP}$$

$$F1\ Score = \frac{2 \times precision \times recall}{precision + recall}$$

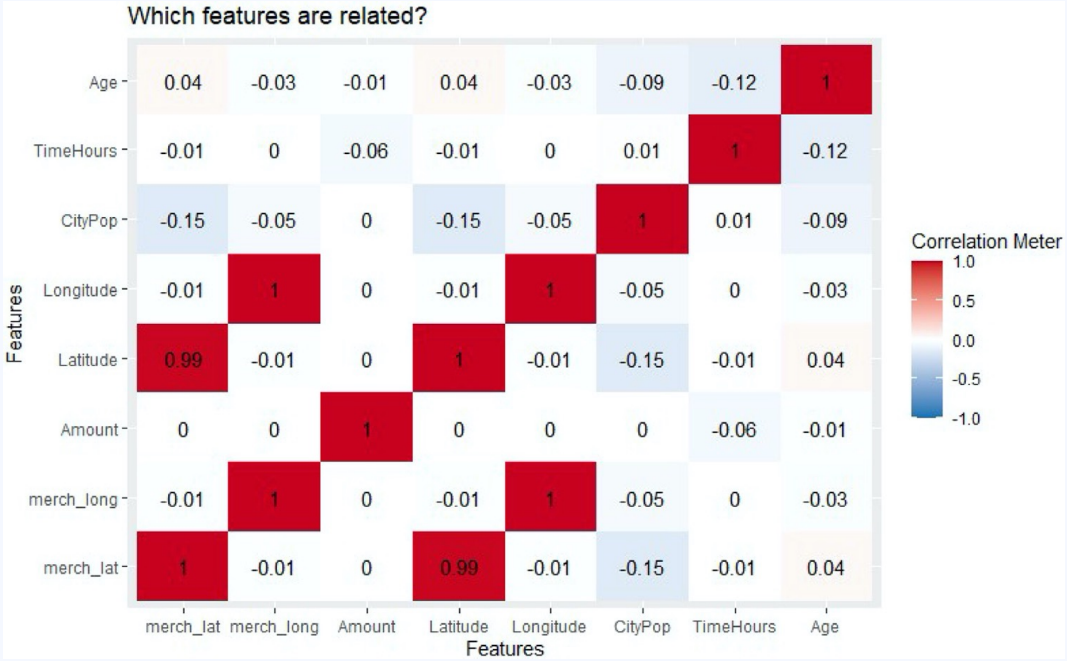
آنالیز داده ها

Table 4
Transaction description.

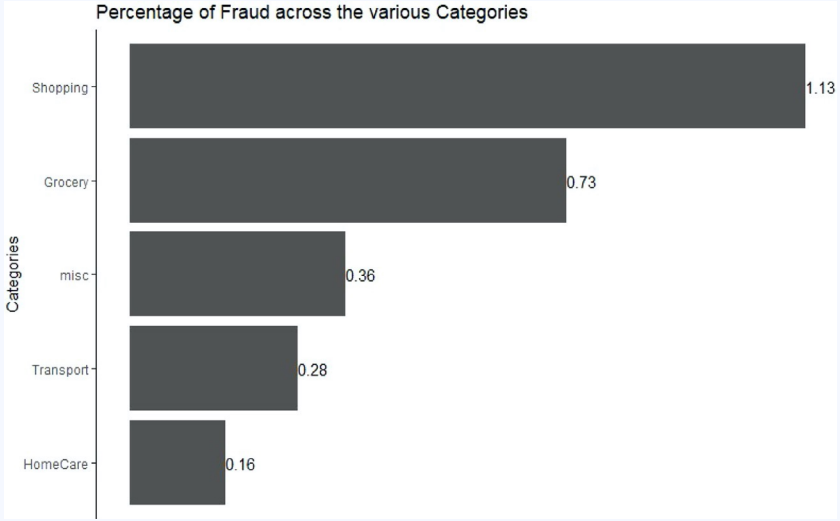
Description	Fraud	Non-Fraud
Total	2135	482 672
Percentage (%)	0.4%	99.6

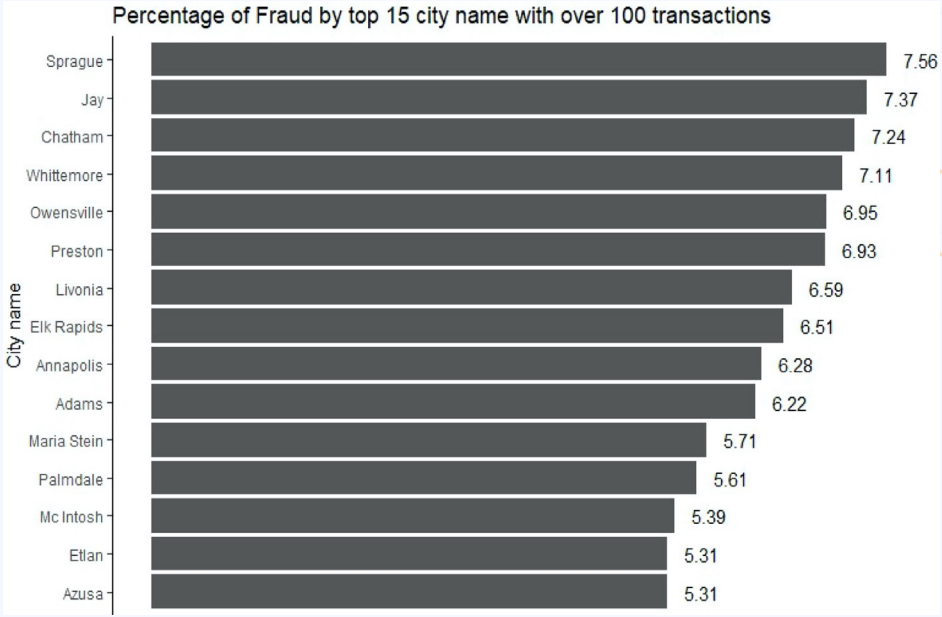


آنالیز داده ها(ادامه)

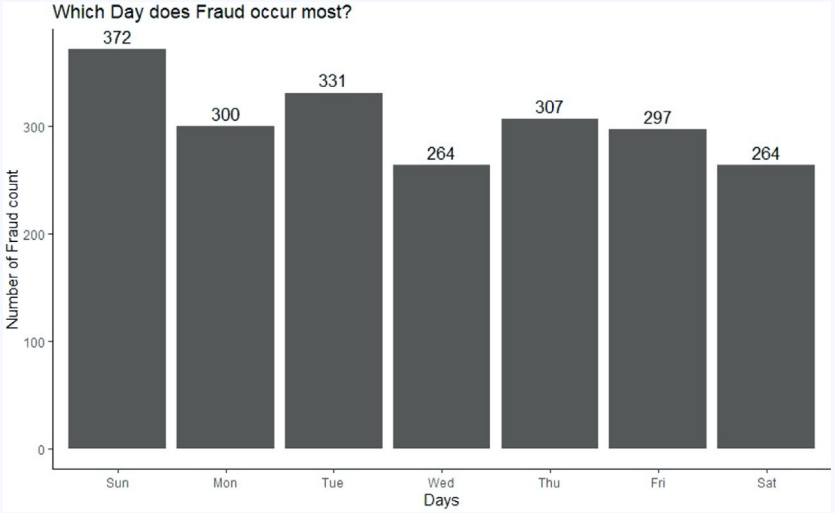


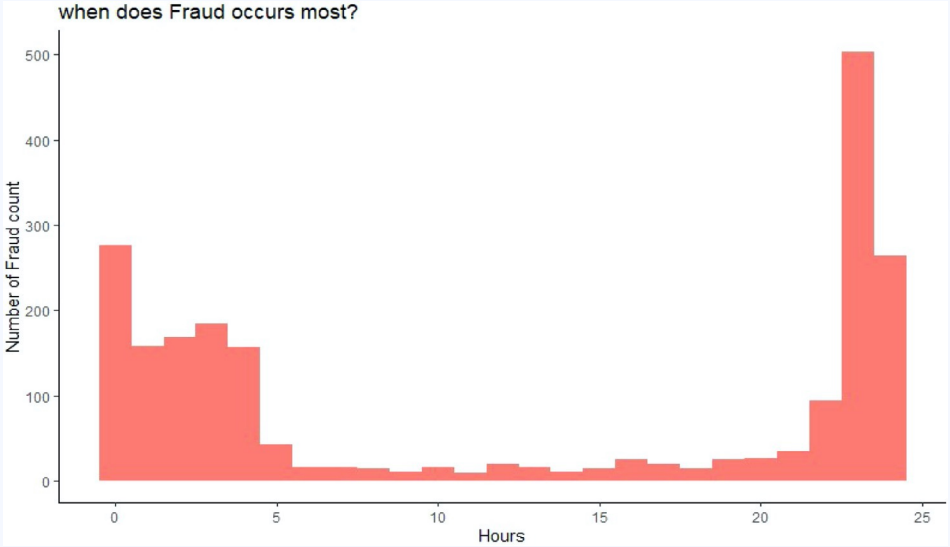
آنالیز داده ها(ادامه)





آنالیز داده ها(ادامه)





نتایج درخت تصمیم

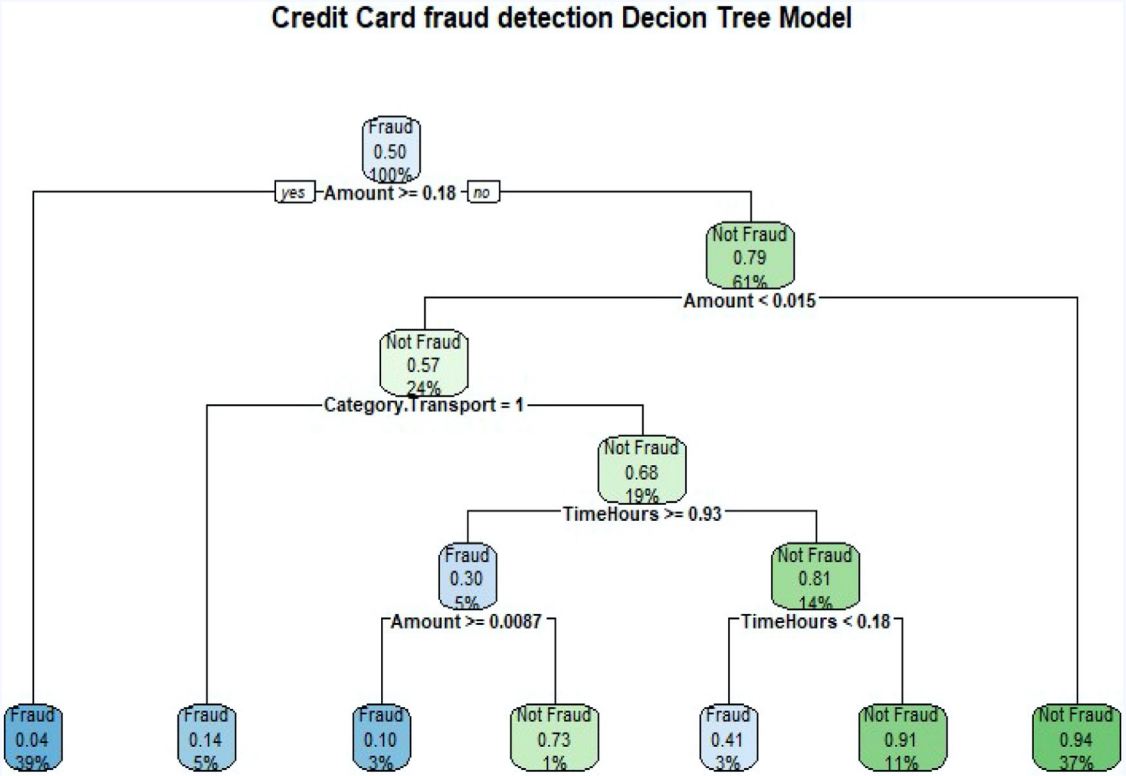
Table 5
Confusion matrix of prediction using decision tree.

	Reference	
Prediction	Fraud	Not Fraud
Fraud	397	8085
Not Fraud	30	88 449

Table 6
Performance of the decision tree algorithm.

Metric measure	Estimate
Accuracy	0.92
Sensitivity	0.93
Specificity	0.92

درخت تصمیم نهایی



نتایج جنگل تصادفی

Table 7
Confusion matrix of prediction using random forest.

	Reference	
Prediction	Fraud	Not Fraud
Fraud	409	4052
Not Fraud	18	92 482

Table 8
Performance of the random forest algorithm.

Metric measure	Estimate
Accuracy	0.96
Sensitivity	0.97
Specificity	0.96

نتایج رگرسیون لجستیک

Table 9
Confusion matrix of prediction using logistics regression.

	Reference	
Prediction	Fraud	Not Fraud
Fraud	325	7731
Not Fraud	102	88 803

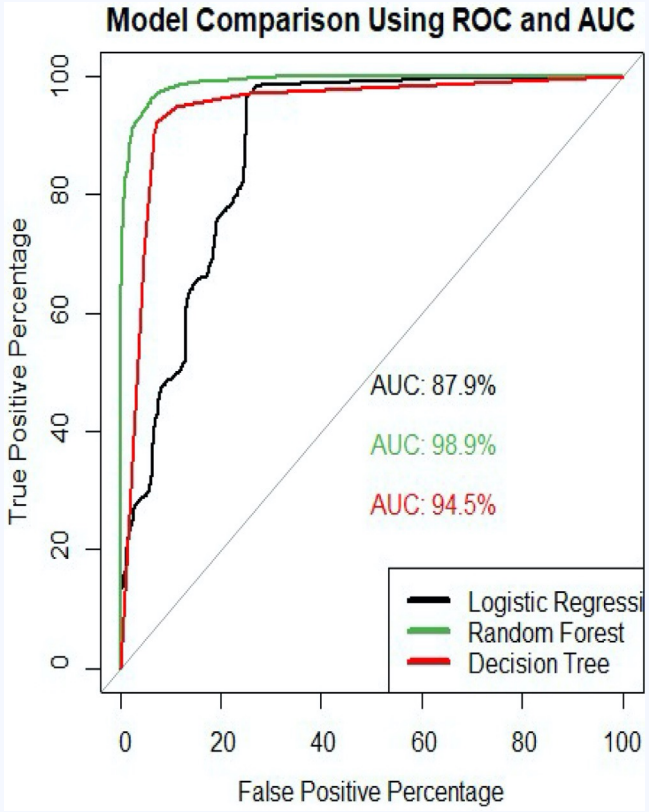
Table 10
Performance of the logistic regression algorithm.

Metric measure	Estimate
Accuracy	0.92
Sensitivity	0.76
Specificity	0.92

مقایسه مدل ها

Table 11
Comparing the models' performances.

Model name	Accuracy	F1-Score	Recall	Precision	Specificity
Decision tree	0.92	0.09	0.93	0.05	0.92
Random forest	0.96	0.17	0.97	0.09	0.96
Logistics regression	0.92	0.08	0.76	0.04	0.92



نتیجه گیری

1. در این مطالعه از سه مدل دسته‌بندی (رگرسیون لجستیک، درخت تصمیم، و جنگل تصادفی) استفاده شده است
2. با استفاده از تکنیک کاهش نمونه , جنگل تصادفی با دقت ۹۶ درصد بهترین عملکرد را ارائه کرده است
3. بر اساس آنالیز دیتا , کلاهبرداری ها بین ساعت ۱۰ شب تا ۵ صبح رخ اتفاق افتاده است
4. بر اساس آنالیز دیتا , افراد بالای ۶۰ سال بیشتر در خطر کلاهبرداری قرار دارند
5. توصیه می شود به افراد مسن خدمات حضوری را در اولویت قرار دهند همچنین بین ساعت ۱۰ شب تا ۵ صبح تدابیر امنیتی را افزایش دهند

برای آینده

1. سایر الگوریتم های یادگیری ماشینی تحت نظارت را می توان در مطالعات آینده با داده های سطح ملی یا بین منطقه ای در نظر گرفت.
2. مطالعه حاضر همچنین می‌تواند در بخش بهداشت و سایر بخش‌ها برای اهداف طبقه‌بندی گسترش یا اعمال شود.

با تشکر

پرسش و پاسخ

ارتباط با من:
mahdimahdiani@ymail.com