چکیده :

در این پژوهش هدف ما ایجاد مدل های معین کلاس بندی از جمله svm, decision trees, random forest و ... و امتحان آن ها بر روی دیتاست کلاهبرداری یا خیر درخواست های بیمه برای تصادفات ماشین ها می باشد. در این جریان ما به بررسی دیتا و سپس بالانس کردن آن پرداختیم و پس از آمادی سازی دیتا برای ترین کردن مدل ، خروجی آن ها را با مدل های مختلف بر اساس stratified cross validation ارائه کردیم. تا بهترین مدل مشخص شود . در این پروسه سعی بر بهبود نتیجه مدل ها داشته ایم.

مقدمه :

از بین مدل های سوپروایزد کلاس بندی کدام مدل می تواند برای دیتاست مشخص مورد پژوهش در این مقاله عملکرد بهتری داشته باشد؟ چه روش هایی برای بهبود عملکرد ه مدل وجود دارد؟ در راستای پاسخ به این سوال های مطرح شده ابتدا به بررسی دیتاست پرداختیم که متشکل از 32 فیچر 1 تارگت و 15420 داده بوده. پس از بررسی ارتباط بین فیچر ها و بررسی بالانس بودن و اعمال سه روش مختلف over sampling , under-sampling و SMOTE و اسکیل کردن های فیچر ها مدل های مختلف مطرح شده را بر روی دیتاست فیت کردیم و سپس خروجی را به stratified cross validation با تشکیل پنج دیتاست مختلف و خروجی گرفتن از آن ها عملکرد را مقایسه کردیم. و سعی در بهبود عملکرد آن ها داشته ایم.

روش ها :

ابتدا کتابخانه های بیسیک کار را لود می کنیم.

Import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

سپس دیتاست vehicle insurance claim detection را در دیتا فریم df لود می کنیم. سپس برای بدست آوردن تعداد فیچر ها و سایز دیتاست از دستور data.shape استفاده می کنیم که خروجی آن بدین شکل می باشد:

(15420, 33)

این یعنی دیتا ست ما شامل 33 فیچر است که یکی از آن ها target روش سوپروایزد و باقی فیچر هایی هستند که مدل ه بر اساس آن ها ترین می شوند. و سایز دیتا ست هم 15420 می باشد. حال نیاز است که بدانیم فیچر های ما چه هستند و تارگت را از بین آن ها تشخیص دهیم. پس از دستور df.columns استفاده می کنیم:

Index(['Month', 'WeekOfMonth', 'DayOfWeek', 'Make', 'AccidentArea',

'DayOfWeekClaimed', 'MonthClaimed', 'WeekOfMonthClaimed', 'Sex',

'MaritalStatus', 'Age', 'Fault', 'PolicyType', 'VehicleCategory',

'VehiclePrice', 'FraudFound\_P', 'PolicyNumber', 'RepNumber',

'Deductible', 'DriverRating', 'Days\_Policy\_Accident',

'Days\_Policy\_Claim', 'PastNumberOfClaims', 'AgeOfVehicle',

'AgeOfPolicyHolder', 'PoliceReportFiled', 'WitnessPresent', 'AgentType',

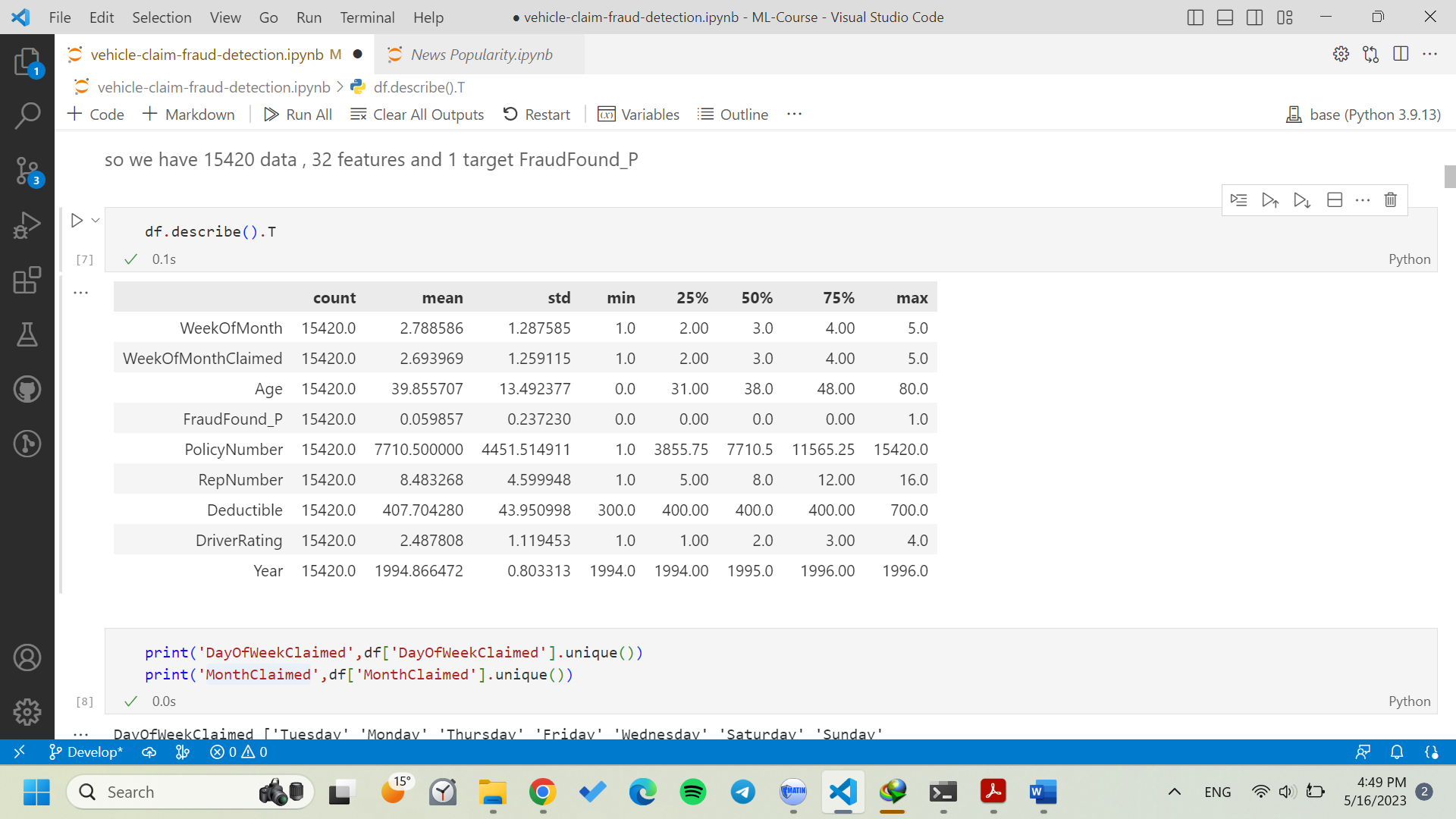
'NumberOfSuppliments', 'AddressChange\_Claim', 'NumberOfCars', 'Year',

'BasePolicy'],

dtype='object')

از بین فیچر های نوشته شده فیچر تارگت همان FraudFound\_P می باشد که خروجی آن احتمالا به شکل yes or no می باشد.

برخی فیچر های ما numeric و برخی از جنس object هستند پس در ادامه کار نیاز داریم که دیتارا encode کنیم که همه آن ها به شکل عددی شوند.



جدول بالا شهود حدودی از توزیع داده های عددی به ما می دهد. تفاوت فاحش رنج اعداد بالا نشان دهنده این است که در ادامه نیاز به اسکیل کردن داده ها داریم. با بررسی مقدار هایی که هر فیچر می تواند داشته باشد نویز را در دو فیچر 'DayOfWeekClaimed'و 'MonthClaimed'شناسایی می کنیم. با دستور:

print('DayOfWeekClaimed',df['DayOfWeekClaimed'].unique())

print('MonthClaimed',df['MonthClaimed'].unique())

که خروجی مقابل را به ما می دهد:

DayOfWeekClaimed ['Tuesday' 'Monday' 'Thursday' 'Friday' 'Wednesday' 'Saturday' 'Sunday'

'0']

MonthClaimed ['Jan' 'Nov' 'Jul' 'Feb' 'Mar' 'Dec' 'Apr' 'Aug' 'May' 'Jun' 'Sep' 'Oct'

'0']

مقدار 0 در این داده ها نویز می باشد که نیاز است حذف شوند پس کد مقابل را اجرا می کنیم:

df **=** df.loc[df['DayOfWeekClaimed']**!=**'0']

df **=** df.loc[df['MonthClaimed']**!=**'0']

و حال برای اینکه ببینیم چه تعداد داده حذف شده اند از دستور df.shape استفاده می کنیم. که خروجی آن به شکل زیر است:

(15419, 33)

این عدد حاکی از این است که فقط یک دیتا بوده که روز هفته و ماه آن صفر وارد شده بود.

برای یافتن missing value ها دستور df.isna().sum() را اجرا می کنیم. که خروجی آن نشان می هد داده گمشده ای وجود ندارد.

Month 0

WeekOfMonth 0

DayOfWeek 0

Make 0

AccidentArea 0

DayOfWeekClaimed 0

MonthClaimed 0

WeekOfMonthClaimed 0

Sex 0

MaritalStatus 0

Age 0

Fault 0

PolicyType 0

VehicleCategory 0

VehiclePrice 0

FraudFound\_P 0

PolicyNumber 0

RepNumber 0

Deductible 0

DriverRating 0

Days\_Policy\_Accident 0

Days\_Policy\_Claim 0

PastNumberOfClaims 0

AgeOfVehicle 0

AgeOfPolicyHolder 0

...

AddressChange\_Claim 0

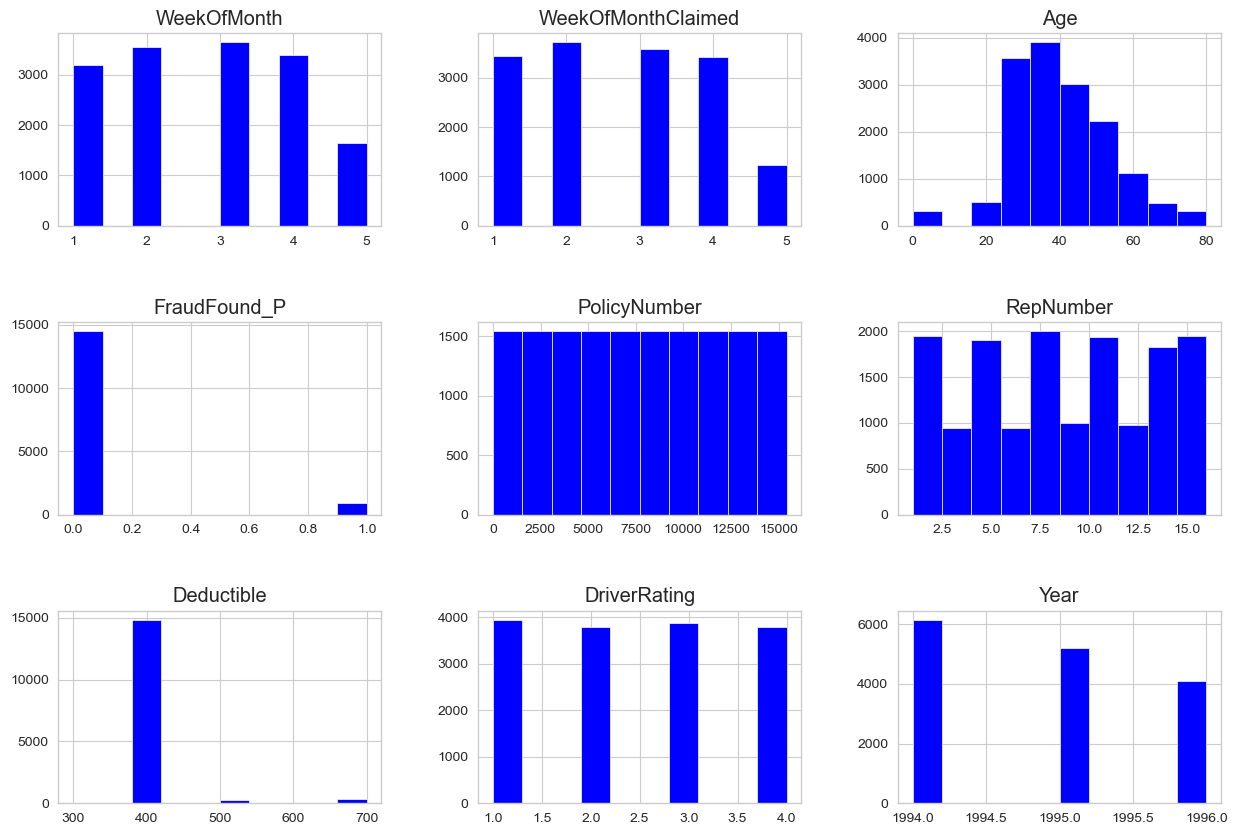
NumberOfCars 0

Year 0

BasePolicy 0

dtype: int64

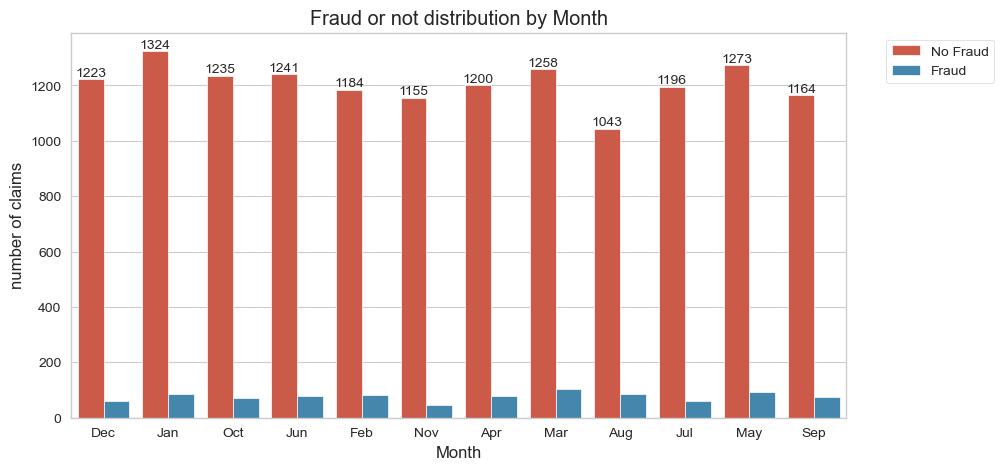
حال برای درک بهتر داده ها به data visualizing می پردازیم.



نمودار های بالا از متغیر های عددی چند چیز را به ما نشان می دهند:

1. نمودار fraudFound\_P نشان دهنده ایمبالانس بودن دیتاست که نیاز به بالانس شدن که ادامه دارد.
2. نمودار age نشان دهنده حضور تعدادی داده پرت با توجه به وجود راننده صفر ساله می باشد.
3. در اواسط ماه میزان درخواست برای بیمه تصادفات بیشترین مقدار است.
4. میزان درخواست بیمه تصادفات در بین راننده هایی با rating های مختلف فرق چندانی ندارد.

نمودار های بالا برحسب کل دیتا و فارغ از کلاهبرداری بودن یا نبودن آن ها بوده . برای درک بهتر تفاوت توزیع فیچر ها در کیس های کلاهبرداری یا غیر کلاهبرداری نمودار بهتری را برای فیچر ها رسم می کنیم.



نمودار بالا این توزیع را در ماه های مختلف رسم می کند. چون با دیتای ایم بالانس مواجه هستیم تعداد کیس های کلاهبرداری در هر ماه بسیار کمتر از کیس های غیرکلاهبرداری می باشد. هرچند در ماه های متفاوت مقدار هردوی این ها تفاوت چندانی ندارد اما به طور کلی بیشترین درخواست بیمه برای تصادفات در ماه های jan, mar, may بنظر بیشتر از سایر ماه ها می آید. برای بررسی نرخ کلاهبرداری بودن درخواست کد ratio مربوطه را محاسبه می کنیم.

که خروجی بدین شکل می باشد:

Dec 0.04824902723735409

Jan 0.06165839829907867

Oct 0.05363984674329502

Jun 0.06056018168054504

Feb 0.06477093206951026

Nov 0.038301415487094086

Apr 0.0625

Mar 0.075

Aug 0.07453416149068323

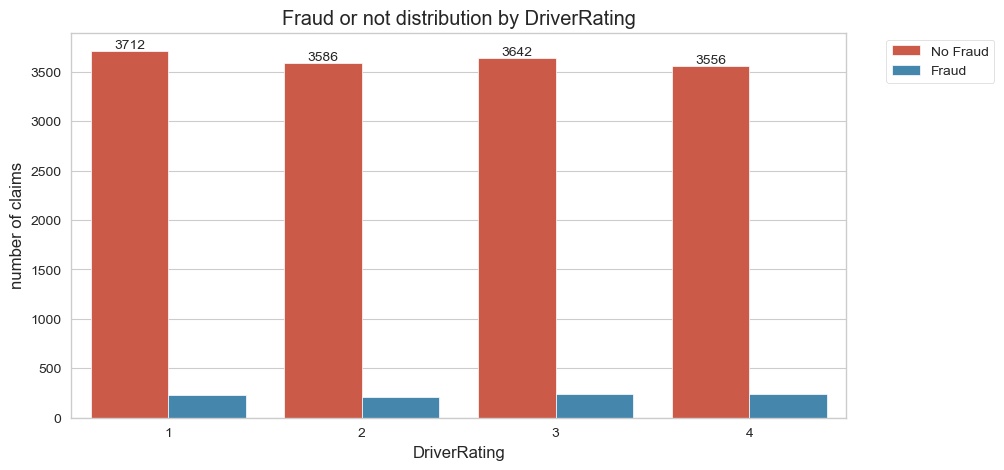
Jul 0.04777070063694268

May 0.06876371616678859

Sep 0.06129032258064516

مشاهده می کنیم ماه های mar , aug دارای بیشترین نرخ کلاهبرداری هستند.

این نمودار را برای فیچر driver rating نیز رسم می کنیم و نرخ کلاهبرداری را برای آن بدست می آوریم.



1 0.058823529411764705

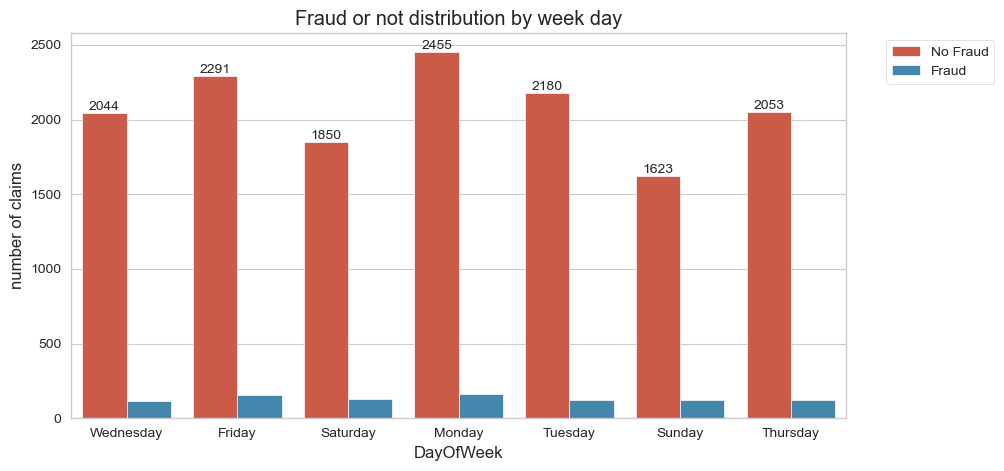
4 0.06198892112898971

3 0.06230690010298661

2 0.05631578947368421

این میزان ratio نشان دهنده این است که کلاهبرداری بودن درخواست بیمه تا حدی با امتیاز راننده رابطه مستقیم دارد.

به ترتیب این پروسه را برای WeekOfDay نیز تکرار می کنیم.

Wednesday 0.053265400648448355

Friday 0.06298568507157465

Saturday 0.06659939455095863

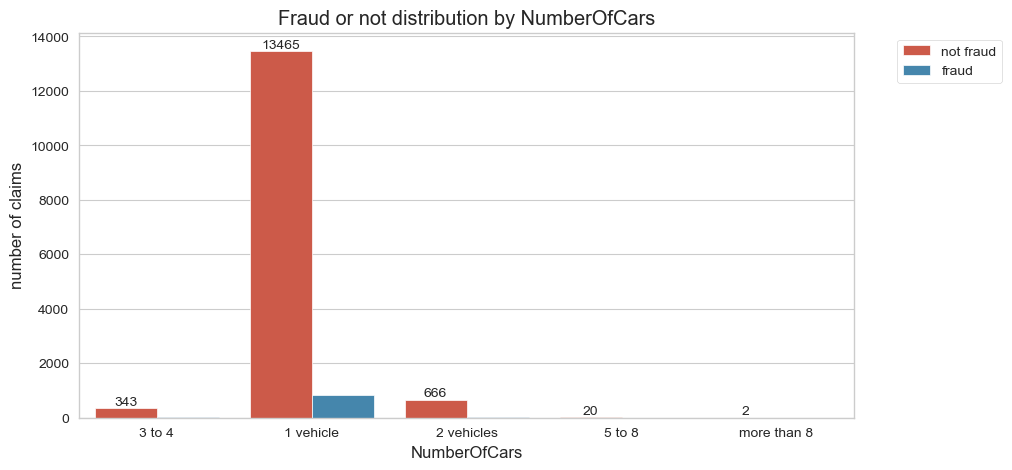
Monday 0.06118546845124283

Tuesday 0.05217391304347826

Sunday 0.06991404011461318

Thursday 0.05522319374137138

مشاهده می شود میزان بیشتر کلاهبرداری اخرهفته که شامل روزهای Sunday و Saturday می باشد ، بیشتر از باقی روزهاست.



3 to 4 0.07795698924731183

1 vehicle 0.059378274537198746

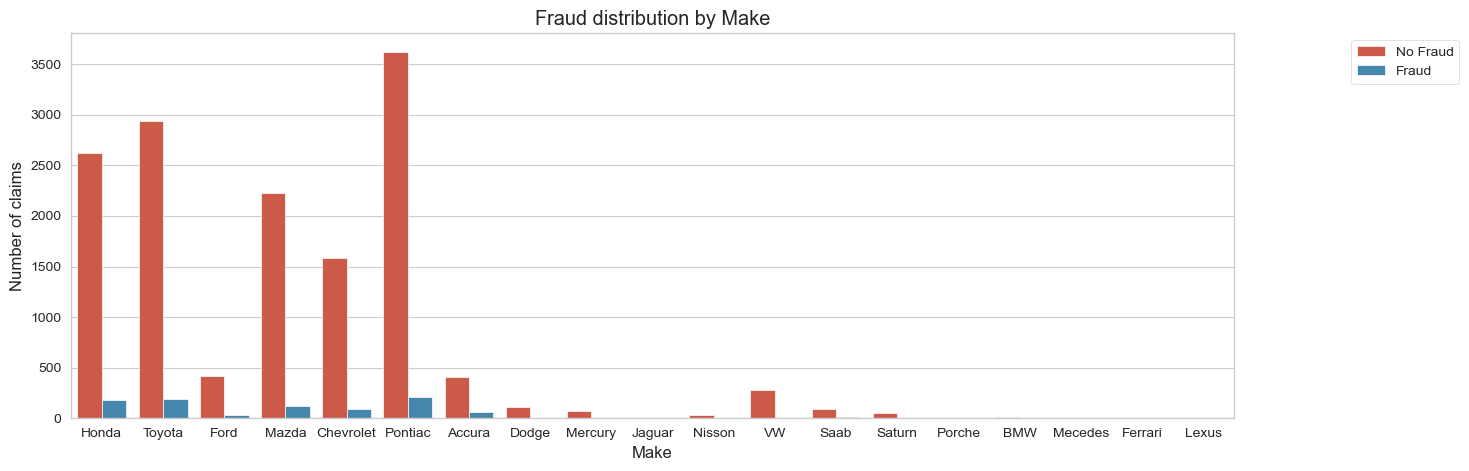
2 vehicles 0.06064880112834979

5 to 8 0.047619047619047616

more than 8 0.0

همانطور که می بینیم بیشترین توزیع درخواست ها در تعداد وسط دارایی ماشین قرار دارد اما نرخ کلاهبرداری افرادی که 3 تا 4 ماشین داشته اند بیشتر از سایرین است.

نمودار و نرخ کلاهبرداری را برای مدل ماشین ها بدست می آوریم:



Honda 0.06392857142857143

Toyota 0.059596283242550466

Ford 0.07333333333333333

Mazda 0.052251486830926085

Chevrolet 0.05591909577632362

Pontiac 0.05551211884284597

Accura 0.125

Dodge 0.01834862385321101

Mercury 0.07228915662650602

Jaguar 0.0

Nisson 0.03333333333333333

VW 0.028268551236749116

Saab 0.10185185185185185

Saturn 0.10344827586206896

Porche 0.0

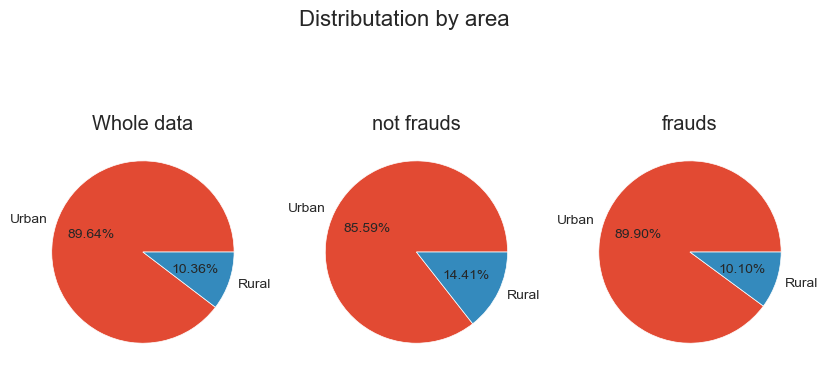
BMW 0.06666666666666667

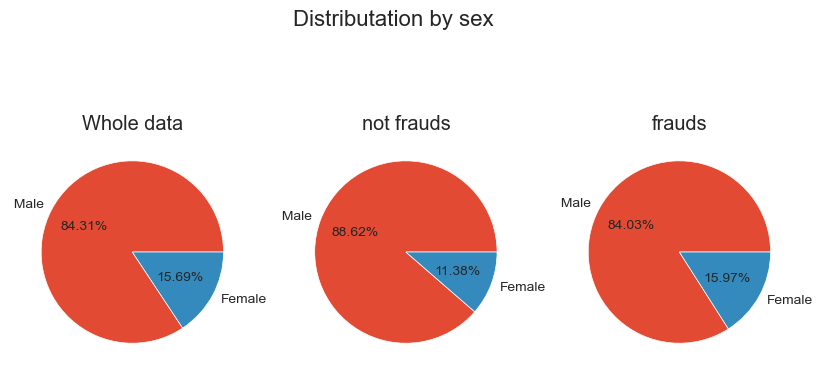
Mecedes 0.25

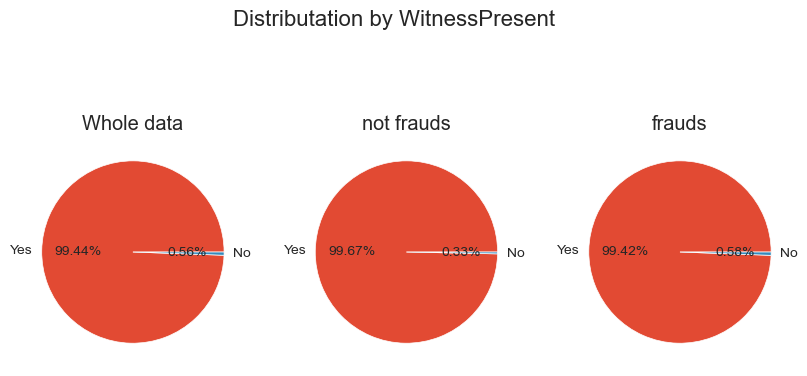
Ferrari 0.0

Lexus 0.0

همانطور که مشاهده می کنیم بیشترین مدل ماشین که دارای ادعای کلاهبرداری بوده با اختلاف مربوط به مدل saab , Saturn می باشد.همچنین با توجه به نمودار در می یابیم که مدل های Honda, Toyota, ford, mazda, Chevrolet, potiac به میزان قابل توجهی درخواست بیمه تصادف بیشتری داشته اند. این می تواند نشانی از مستعدتر بودن این مدل ها برای تصادف باشند.



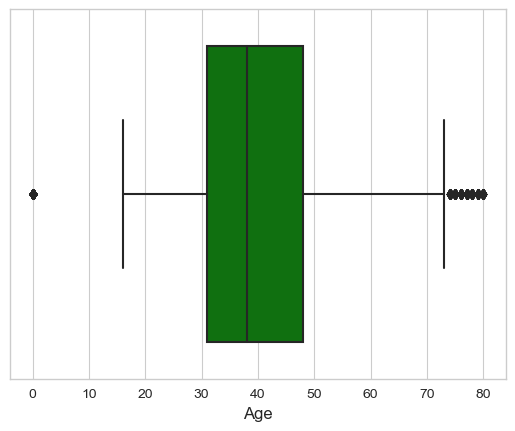




سه نمودار بالا نشان دهنده این توزیع در سه فیچر sex, witnessPresent , area می باشد. به دلیل اینکه این سه فیچر تنها مقادیر باینری اتخاذ می کنند استفاده از نمودار دایره ای انتخاب بهتری برای آنان است.

با مقایسه درخواست ها متوجه میشویم نرخ کلاهبرداری در نقاط شهری ، در جنس زنان و در عدم حضور شاهد بیشتر از حالت دیگر است.

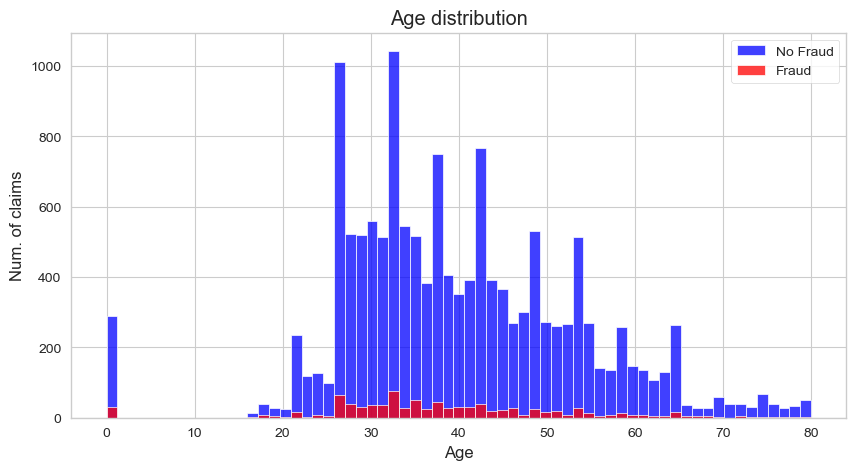
جال پس از مصورسازی دیتا برای یافتن داده های پرتی که در age شناسایی کردیم از boxplot استفاده می کنیم.

نمودار نشان دهنده این است که دارای تعدادی داده پرت در نقطه صفر و

بزرگ تر از 72 می باشیم اما به دلیل پیوسته بودن مقادیر بالا 72 از

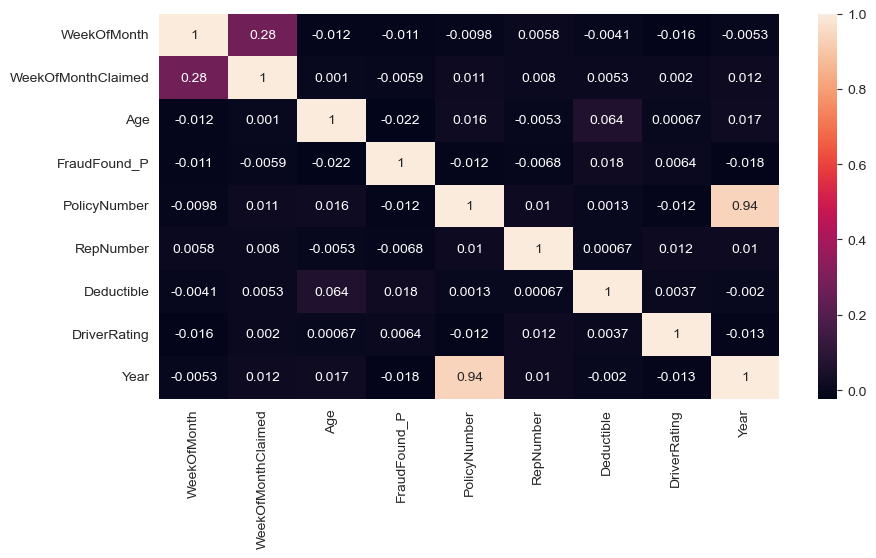
حذف آن ها صرف نظر می کنیم. نمدار histogram فیچر age هم

نشانی بر تایید این امر می باشد که در زیر مصور شده است:



این نمودار تاییدیه بر حذف عنصر پرت age = 0 از دیتاست می باشد.

چیزی که در بخش پیش از مدل سازی باقی مانده بررسی این است که آیا فیپر های ما با هم correlation دارند. یا خیر و برای آزمایش این مورد دیتا را مصور سازی می کنیم:



این نمودار حاکی از این است که میزان قابل توجهی کورلیشن بین فیچر های

PolicyNumber , Year- weekOfMonthClaimed , weekOfMonth و deduction, age وجود دارد که نیاز است به طور اتومات توسط مدل ها یا با اعمال متد فیچر سلکشنی اصلاح شوند.

حال برای آماده سازی دیتا قبل از ترین مدل ها نیاز است داده های غیرعددی که پیش تر بیان شد به طور عددی کد گذاری شوند. این فیچر ها به صورت زیر می باشند:

Index(['Month', 'DayOfWeek', 'Make', 'AccidentArea', 'DayOfWeekClaimed',

'MonthClaimed', 'Sex', 'MaritalStatus', 'Fault', 'PolicyType',

'VehicleCategory', 'VehiclePrice', 'Days\_Policy\_Accident',

'Days\_Policy\_Claim', 'PastNumberOfClaims', 'AgeOfVehicle',

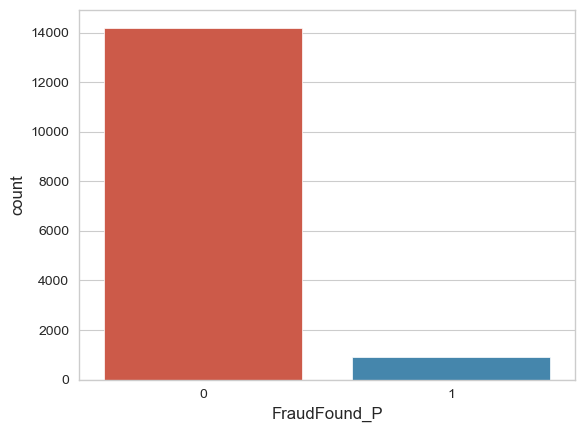
'AgeOfPolicyHolder', 'PoliceReportFiled', 'WitnessPresent', 'AgentType',

'NumberOfSuppliments', 'AddressChange\_Claim', 'NumberOfCars',

'BasePolicy'],

dtype='object')

نیاز است که تمامی حالت های این فیچر ها به اولویت ترتیبی کدگذاری شوند. پس از طی کردن مسیر فهمیدن دیتا و حذف نویز از آن نیاز است اصلاحاتی مانند بالانس کردن ، اسکیل کردن و فیچر سلکشن بر روی دیتا اعمال شود.

با چک کردن وضعیت ایم بالانسی دیتا شروع می کنیم:

0 14208

1 892

Name: FraudFound\_P, dtype: int64

ارقام و نمودار مقابل به ایم بالانس بودن دیتا اشاره دارند.

از جمله روش هایی که برای حل این مشکل می توان به کار برد :

Over sampling, under sampling , SMOTE می باشد

که روش اورسمپلینگ به دلیل زیاد کردن دیتا بر اساس بخش هایی از آن

ممکن است دچار اورفیت و روش اندرسمپلینگ به دلیل کاهش حجم دیتا

ممکن است دچار اندرفیت شود پس در ابتدا روشی که حدس می زنیم

عملکرد بهتری نشان دهد یعنی SMOTE را اعمال می کنیم و سپس

خروجی را با دو تکنیک دیگر هم آزمایش می کنیم تا به مقایسه آن ها بپردازیم.