به نام خدا

پردیس دانشکده های فنی دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

گزارش پروژه چهارم درس هوش مصنوعی

عنوان پروژه : یادگیری ماشین

استاد : سرکار خانم دکتر فدایی

رضوان بهمنی

810197473

مقدمه

در این پروژه یادگیری ماشـــین که در دســـته یادگیری نظارت شـــده(Supervised Learning) قرار می گیرد، درصـــدد آن هستیم که مدلی را آموزش دهیم که توانایی تخمین قیمت یک خودرو را داشته باشد. در این پروژه ابتدا و در فاز اول به کمک چندین تخمین‌گر، مدل هایی را پیاده ســـازی خواهیم کرد و آن ها را بهینه خواهیم کرد. در ادامه و در فاز دوم پروژه، به کمک روش یادگیری گروهی(Ensemble Learning)، مدل های ســـاخته شـــده در فاز اول را ترکیب کرده تا نتایج بهتری بدســـت آوریم. . در ادامه مفصلا مراحل کار برای ساخت و ارزیابی این مدل تشریح شده است.

راهبرد حل مساله

روند اجرایی این پروژه به این صـورت می باشـد که در ابتدا باید داده ورودی مدل آماده شـده و برای آموزش مدل آماده گردد. این داده خام پیش از ورود به روند مدل ســـازی، باید از مرحله پیش پردازش(Pre-Processing) عبور کند. در این مرحله از روش هایی برای مرتب سازی و پاک سازی داده استفاده می شود. پس از آنکه داده غیرقابل آموزش و غیرقابل استفاده به داده قابل آموزش تبدیل می شود، آن را به عنوان ورودی به مدل های یادگیری ماشین می دهیم.

بعد از آماده ســـازی داده در مرحله پیش پردازش، آن را به دو بخش تقســـیم میکنیم. بخش اول براش آموزش مدل(Train) و بخش دوم برای ارزیابی مدل(Test) استفاده می گردد. در این پروژه از 80 درصد داده برای آموزش و از 20 درصد داده برای ارزیابی استفاده شده است. پس از آموزش مدل و ارزیابی آن، شاخص های ارزیابی محاسبه شده و ارائه می شوند.

پرسش 1:

شاخص صرفه اطلاعاتی(Information Gain)

در این قسمت شاخص صرفه اطلاعاتی برای هر یک از ویژگی ها محاسبه شده. نمودار مربوطه در فایل نوت بوک قابل مشاهده است. این نمودار ها، میزان ارتباط فیچر ها با تارگت را بصورت عددی نشان می دهند. هرچه فیچری مقدار بیشتری داشته باشد، همبستگی بیشتری بین آن فیچر و تارگت وجود دارد. به کمک آن می توان میزان مفید بودن فیچر های مختلف در ساخت مدل را بررسی نمود.

پیش پردازه داده

همانطور که پیشتر توضیح داده شد، در این مرحله در صدد آن هستیم که به کمک استفاده از روش هایی بتوانیم داده خام را به بهترین نحو به شـــکل داده قابل آموزش تبدیل کنیم. در این پروژه این فرآیند برای دو ســـتون از ویژگی ها(Features) صورت گرفته است که در ادامه توضیح داده شده است. هم چنین از تبدیل مقیاس(Scale) نیز استفاده شده است.

- داده تاریخ (created_at)

داده تاریخ با فرمت روز-ساعت (day – time) داده شده است. این فرمت قابل آموزش و پردازش نیست. لذا در ابتدا باید فرمت این داده اصلاح شود به این شکل که این داده به دو داده مجزای روز و ساعت تبدیل میشود. برای تبدیل دیتای ساعت، از فرمت 24 ساعته استفاده شده است.

- تبدیل داده های طبقه ای

در این پروژه فیچر های brand و category از نوع داده طبقه ای(Categorical) داده شده است. این نوع داده توسط مدل های یادگیری ماشـــین قابل پردازش نیســـت. لذا نیاز به تبدیل این نوع داده به داده قابل پردازش هســـتیم. برای این تبدیل می توانیم از دو روش Label-Encoding یا One-Hot-Encoding استفاده کنیم. در این پروژه برای مدل هایی که با داده ها به شکل غیرعددی برخورد می کنند، از روش Label-Encoding و برای مدل هایی که با داده ها به شکل عددی برخورد می کنند، از روش One-Hot-Encoding استفاده کرده ایم. در ادامه این دو روش توضیح داده است.

- داده کتگوری (Category)

داده کتگوری بصورت داده طبقه ای(Categorical) داده شده است. این نوع داده توسط مدل های یادگیری ماشین قابل پردازش نیســـت. لذا نیاز به تبدیل این نوع داده به داده قابل پردازش هســـتیم. برای این تبدیل می توانیم از دو روش Label-Encoding یا One-Hot-Encoding اســـتفاده کنیم. در این پروژه برای مدل هایی که با داده ها به شـــکل

غیرعددی برخورد می کنند، از روش Label-Encoding و برای مدل هایی که با داده ها به شـــکل عددی برخورد می کنند، از روش One-Hot-Encoding استفاده کرده ایم. در ادامه این دو روش توضیح داده است.

- تبدیل مقیاس(Scale)

از آنجا که دامنه(Range) مقادیر برای هر یک از ویژگی ها(Features) متفاوت اسـت، ممکن اسـت میزان تاثیر ویژگی ها در متغییر هدف تفاوت ایجاد کند. این مو ضوع در برخی از مدل های یادگیری ما شین رخ می دهد. برای جلوگیری از پیش آمدن این مشـــکل، از تبدیل مقیاس داده ها اســـتفاده می شـــود. در پروژه از مقیاس (1-0) برای تمامی ویژگی ها استفاده شده است تا تاثیر آن ها در متغییر هدف تسطیح شود. البته می توان بطور خاص برای ویژگی هایی، این مقیاس را افزایش یا کاهش داد تا تاثیر آن ها را کنترل کرد. از این امکان در این پروژه صرف نظر شده است.

- حذف دادگان پرت(Outliers Elimination)

میدانیم که دادگان پرت تاثیر منفی در مدل های یادگیری ما شین می گذارند و بهتر ا ست از دادگان حذف گردند. در این پروژه از دو روش مختلف برای حذف دادگان پرت استفاده شده است. در روش اول، بر اساس برازش منحنی نرمال بر روی دادگان هر یک از ویژگی ها، دادگان پرت شنا سایی شده و حذف می گردد. در روش دوم، داده های هر ستون از دادگان ابتدا بر اساس مقادیر مرتب می شوند و سپس درصدی از داده ها با کمترین مقدار و بیشترین مقدار حذف می گردند. هر دو این روش ها تاثیر مثبتی در بهبود نتایج ارزیابی مدل ها داشته است.

پرسش 2

این نوع داده تو سط مدل های یادگیری ما شین قابل پردازش نی ست. لذا نیاز به تبدیل این نوع داده به داده قابل پردازش هســـتیم. برای این تبدیل می توانیم از دو روش Label-Encoding یا One-Hot-Encoding اســـتفاده کنیم. در این پروژه برای مدل هایی که با داده ها به شـــکل غیرعددی برخورد می کنند، از روش Label-Encoding و برای مدل هایی که با داده ها به شکل عددی برخورد می کنند، از روش One-Hot-Encoding استفاده کرده ایم. در ادامه این دو روش توضیح داده است.

- Label-Encoding

در این روش، به هر یک از مقادیری که داده های طبقه ای دارند، عددی صحیح یکتایی اختصاص می یابد. برای مثال اگر متغییر طبقه ای "جنسیت" شامل دو مقدار "مرد" و "زن" باشد، با استفاده از این روش می توان مقادیر "0" و "1" را بترتیب برای مقادیر مذکور در نظر بگیریم. یکی از اشـــکالات این روش برای مدل هایی اســـت که با داده ها بعنوان متغییر های عددی رفتار می کنند. این مدل ها برای مثال فوق، ارزش بیشـــتری به مقدار "1" می دهند در صـــورتی که این امر مطلوب نیست چرا که این دو مقدار تفاوتی از نظر ارزشی ندارند.

- One-Hot-Encoding

در این روش، به ازای هر یک از مقادیری که یک متغییر دسته ای دارد، ستون جدیدی از داده اضافه می شود. حال به ازای هر ردیف از داده ها، مقداری که متغییر دسته ای دارد، برای ستون مختص آن مقدار، مقدار "1" و برای بقیه ستون ها مقدار "0" می گیرد. در این روش، به تعداد مقادیری که یک متغییر دسته ای دارد، متغییر جدید به دیتاست اضافه می شود. اگر تعداد این مقادیر زیاد باشد، تعداد زیادی ستون جدید به دیتاست اضافه می شود که باعث شلوغ شدن دیتاست و کند شدن پردازش در دیتاست های بزرگ می شود. مزیت مهم این روش نسبت به روش Label-Encoding این است که مقادیر "0" و "1" دارای ارزش عددی هستند.

مقایسه دو روش Label-Encoding و One-Hot-Encoding در قالب یک مثال در شکل زیر نشان داده شده است.

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

من برای شاخصه‌ی کتگوری از روش Label Encoding و برای شاخصه Brand از روش One Hot Encoding استفاده کردم.

پرسش 3:

داده های موجود در ستون های title و description برای آن که قابل پردازش شوند از این روش استفاده میکنیم که تعداد تکرار های هر واژه در هر ردیف را بدست می اوریم و عدد، یک معیار برای Sample آن ردیف خواهد شد.

این کار دقت مدل را افزایش میدهد چرا که یکی از معیارهای اصلی در تعیین target توضیحات نوشته شده در این 2 ستون است.

پرسش 4:

راه های مقابله با داده های گزارش نشده(missing values) را می توان در دو دسته کلی قرار دارد.

- حذف ردیف هایی که دارای داده های گزارش نشده هستند.

در این روش تمامی ردیف هایی که دارای ستونی با داده گزارش نشده هستند، از دیتاست حذف می شوند. این روش برا; دیتاست هایی که تعداد دادگان کمی دارند مناسب نیست.

- جایگزین کردن داده های گزارش نشده با مقادیر دیگر

در این روش تمام دادگان گزارش نشده با مقادیر دیگری جایگزین خواهند شد. این مقادیر خود می توانند از روش ها; مختلفی بدست آیند. از جمله این روش ها میتوان موارد زیر را نام برد:

- یک مقدار متفاوت از مقادیر دیگر ردیف ها
- یک مقدار که بصورت رندوم از ردیف های دیگر انتخاب شده باشد.
- میانگین، میانه و یا مد دادگان آن ستون
- مقداری که به کمک مدل های پیش بینی بدست آید.

بطوری کلی نمی توان یک روش را از بقیه روش ها مناسب تر دانست و این برتری بین روش ها در هر مسئله متفاوت است.

در این مساله داده های NAN حذف شدند.

پرسش 5:

تقسیم دادگان

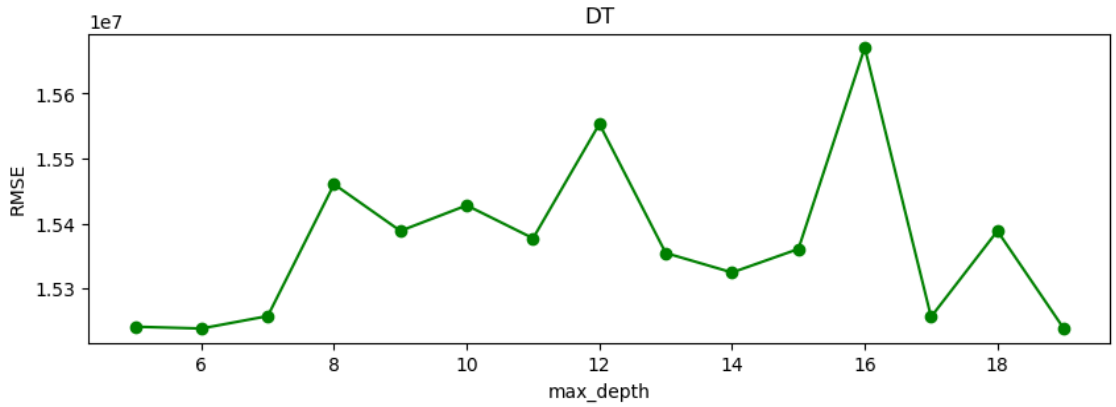
دادگان پروژه باید به دو دسته برای آموزش(Train) و ارزیابی(Test) تبدیل گردد. از دسته آموزش برای آموزش مدل های یاد گیری ماشین استفاده می شود. در ادامه به کمک داده های ارزیابی، مدل ارزیابی می شود و پارامتر های ارزیابی مدل محاسبه می شوند. شایان ذکر است که داده آموزش و داده ارزیابی با یکدیگر هم پوشانی ندارند و برای ارزیابی مدل از داده ای استفاده می شود که مدل تا به حال آن را ندیده است. در این پروژه از 80 درصد دادگان برای آموزش و از 20 درصد مابقی برای ارزیابی استفاده شده است.

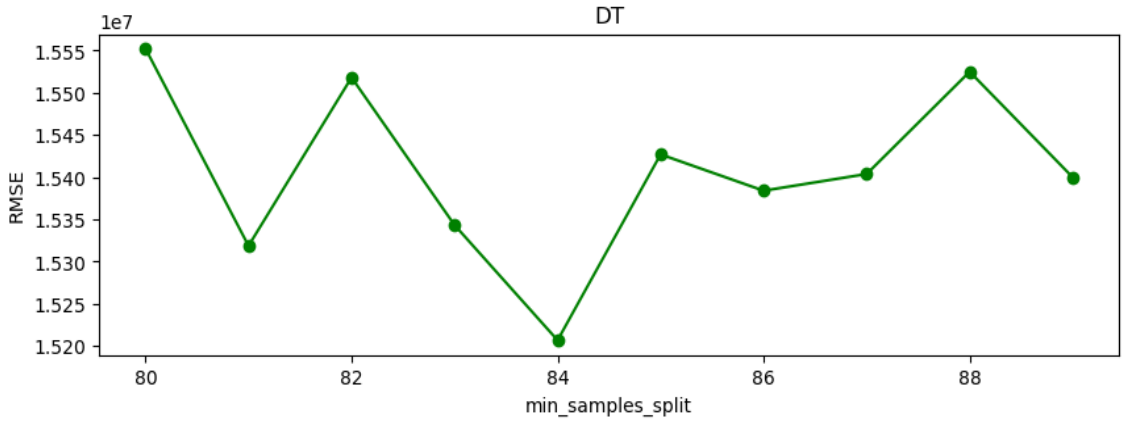
اگر سهم داده های آموزش را کاهش دهیم، آنگاه مدل جامعیت خود را از دست می دهد زیرا روی دادگان کمتری آموزش دیده است و احتمال Overfitting در آن مدل بیشتر می شود.

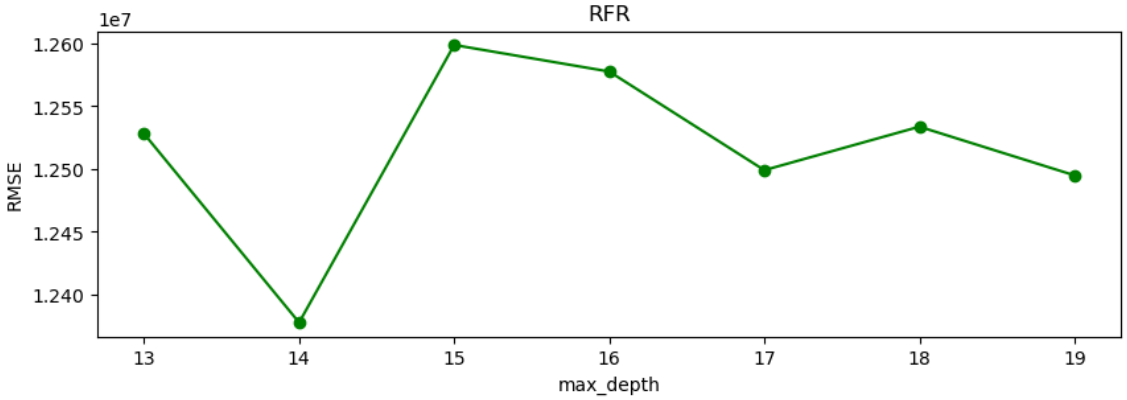
اگر سهم داده های آموزش را خیلی زیاد کنیم، آنگاه داده های کمی برای ارزیابی مدل خواهیم داشت ک باعث می شود نتوان به شکل صحیح مدل را ارزیابی کرد. دراین حالت ممکن است Underfitting اتفاق افتاده باشد اما به دلیل حجم محدود داده های تست، نتوانیم این مشکل را تشخیص دهیم!

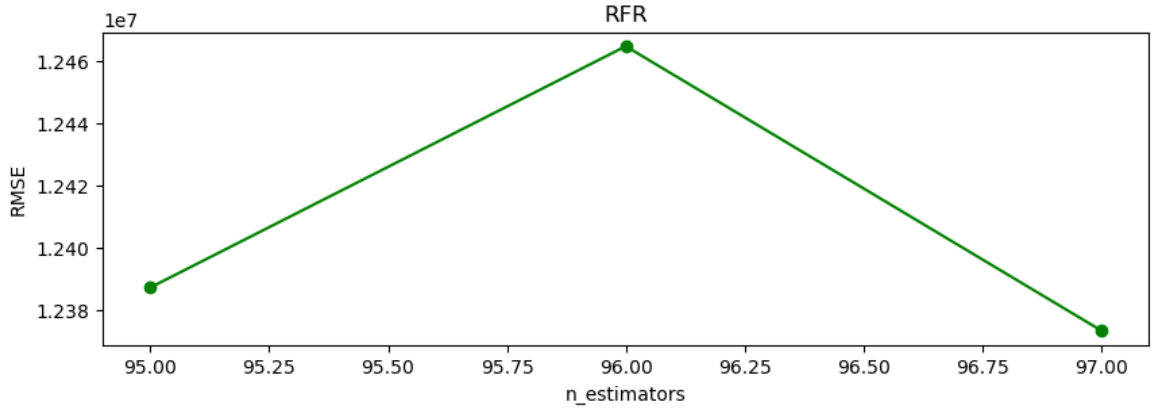
پرسش 6:

اگر Max_depth را افزایش دهیم، احتمالا رخ دادن Overfitting بیشتر می شود. اگر مقدار آن را کم کنیم، قدرت یادگیری مدل کمتر می شود و Underfitting رخ می دهد. لذا باید این پارامتر tune شود و مقدار بهینه آن انتخاب شود.


DT
1e7
1.56
1.55
1.54
1.53
RMSE
6
8
10
12
14
16
18
max_depth


DT
1e7
1.555
1.550
1.545
1.540
1.535
1.530
1.525
1.520
RMSE
80
82
84
86
88
min_samples_split


RFR
1e7
1.260
1.255
1.250
1.245
1.240
RMSE
13
14
15
16
17
18
19
max_depth


RFR
1e7
1.246
1.244
1.242
1.240
1.238
RMSE
95.00
95.25
95.50
95.75
96.00
96.25
96.50
96.75
97.00
n_estimators

یادگیری گروهی(Ensemble Learning)

در این مرحله درصدد آن هستیم که با استفاده از روش های یادگیری گروهی، دقت مدل های یادگیری ماشین پیاده سازی شده در مرحله قبل را افزایش دهیم.

- روش Bagging

در این روش تخمین‌گرهایی از مدل های یکسان ساخته می شوند. به هر کدام از این دسته بند ها پارامتر های اصلی یکسانی داده می شود. اما هر کدام از دسته بند ها دارای ردیف ها و ستون های متفاوتی از داده ها هستند. این روش کمک می کند تا هر کدام از دسته بند ها به شکل متفاوتی با دادگان روبرو شوند. در نهایت از بین دسته هایی که هر کدام از دسته بند ها پیش بینی کرده اند، دسته با بیشترین تکرار انتخاب می شود.

پرسش 7:

پارامتر واریانس نشان دهنده آن است که مدل چقدر general است و تا چه حد بر روی داده های جدید می تواند جامعیت خود را حفظ کند. هر چه میزان واریانس مدل کمتر باشد، بر روی داده های جدید عملکرد بهتری نشان می دهد.

پارامتر Bias نشان دهنده قدرت یادگیری مدل روی داده آموزش است. هرچه میزان Bias بیشتر شود، یادگیری مدل کمتر می شود.

اگر Bias زیاد باشد و واریانس کم باشد، Underfitting رخ می دهد و مدل حتی بر روی داده های آموزش فیت نمی شود.

اگر Bias کم باشد و واریانس زیاد باشد، Overfitting رخ می دهد و مدل جامعیت خود را از دست داده و نویز ها را نیز یاد می گیرد و روی دیتای جدید عملکرد خوبی نشان نمی دهد.

پیش بینی می شود که هر دو پارامتر Bias و واریانس در مدل های Random Forest مقدار کمتری نسبت به مدل های Decision Tree داشته باشند.

RMSE for DT : 15262659.20468844
MSE for DT : 232948765998468.16

RMSE for RFR : 12432035.248538334
MSE for RFR : 154555500420899.2