دانشگاه خوارزمی تهران

دانشکده فنی مهندسی

مهندسی کامپیوتر

گزارش ارائه

Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI

نگارش

سامان عطارکاشانی

استاد درس

دکتر منصوری

بهمن ۱۴۰۱

فهرست مطالب

صفحه

فصل اول مقدمه..... ۱
۱-۱- مقدمه..... ۱
۱-۲- اهمیت موضوع..... ۲
۱-۳- مخاطب و هدف توضیح..... ۳
۱-۴- محبوبیت مشارکت XAI..... ۴
فصل دوم توضیح اصطلاحات و لغات تخصصی..... ۵
۲-۱- مقدمه..... ۶
۲-۲- شفافیت..... ۷
۲-۳- تعاریف..... ۸
فصل سوم XAI در مدل‌های یادگیری ماشین..... ۱۰
۳-۱- مقدمه..... ۱۱
۳-۲- رگرسیون خطی/ لجستیک..... ۱۲
۳-۳- درخت تصمیم گیری..... ۱۳
۳-۴- K نزدیکترین همسایه..... ۱۳
۳-۵- مدل های مبتنی بر قانون..... ۱۳
۳-۶- مدل های بیز..... ۱۴
۳-۷- روش‌های افزایش توضیح‌پذیری..... ۱۵
۳-۸- مدل های ترکیبی شفاف و جعبه سیاه..... ۱۷
فصل چهارم فرصت‌های تحقیق و هوش مصنوعی مسئول..... ۱۸
۴-۱- فرصت های تحقیق در XAI..... ۱۹
۴-۲- هوش مصنوعی مسئول..... ۱۹
فصل پنجم چرا باید به تو اعتماد کنم؟ ۲۰
مدل توضیح پذیر LIME..... ۲۱
فصل ششم نتیجه‌گیری و جمع بندی..... ۲۳
فصل هفتم کد و تفسیر آن به صورت گام به گام..... ۲۵
منابع و مراجع..... ۳۳

فهرست شکل‌ها

صفحه

شکل ۱ – مثالی از XAI ۲
شکل ۲ – نمودار مخاطب توضیح‌پذیری و چرایی ۴
شکل ۳ – نمودار افزایش مشارکت با کلیدواژه‌های مرتبط با XAI از سال ۲۰۱۲ تا ۲۰۱۹ ۴
شکل ۴ – سطوح شفافیت ۷
شکل ۵ – رابطه‌ی اهداف و مخاطب هدف ۹
شکل ۶ – تقسیم‌بندی مدل‌ها از نظر شفافیت ۱۱
شکل ۷ – تقسیم‌بندی مدل‌های یادگیری ماشین و سطوح شفافیت ۱۴
شکل ۸ – روش‌های توضیح‌پذیری ۱۵
شکل ۹ – استفاده از چارچوب LIME ۱۷
شکل ۱۰ – نمودار مدل ترکیبی ۱۷
شکل ۱۱ – نمودار هوش مصنوعی مسئول ۱۹
شکل ۱۲- شمای کلی از مدل‌های توضیح پذیر ۲۲

فصل اول

مقدمه

۱-۱- مقدمه

هوش مصنوعی هسته‌ی اصلی بسیاری از بخش‌های فعالیتی است که فناوری‌های اطلاعاتی جدید را پذیرفته‌اند. در حالی که ریشه‌های هوش مصنوعی به چندین دهه قبل بازمی‌گردد، اجماع واضحی در مورد اهمیت فوق‌العاده‌ای که امروزه توسط ماشین‌های هوشمند دارای قابلیت‌های یادگیری، استدلال و سازگاری مشخص می‌شود، وجود دارد. به واسطه این قابلیت‌ها است که روش‌های هوش مصنوعی هنگام یادگیری حل وظایف محاسباتی پیچیده‌تر، به سطوح بی‌سابقه‌ای از عملکرد دست می‌یابند و آنها را برای توسعه آینده جامعه انسانی محوری می‌سازند . پیچیدگی سیستم های مجهز به هوش مصنوعی اخیراً به حدی افزایش یافته است که تقریباً برای طراحی و استقرار آنها نیازی به دخالت انسانی نیست. هنگامی که تصمیمات ناشی از چنین سیستم‌هایی در نهایت بر زندگی انسان‌ها تأثیر می‌گذارد (مثلاً در پزشکی، قانون یا دفاع)، نیاز آشکاری برای درک چگونگی ارائه چنین تصمیم‌هایی توسط روش هوش مصنوعی وجود دارد.

در حالی که اولین سیستم های هوش مصنوعی به راحتی قابل تفسیر بودند، در سال های گذشته شاهد ظهور سیستم های تصمیم گیری مبهم مانند شبکه های عصبی عمیق بوده ایم. موفقیت تجربی مدل های یادگیری عمیق مانند DNN ها از ترکیبی از الگوریتم های یادگیری کارآمد و فضای پارامتریک عظیم آنها ناشی می‌شود. فضای دوم شامل صدها لایه و میلیون ها پارامتر است که باعث می‌شود DNN ها به عنوان مدل های جعبه سیاه پیچیده در نظر گرفته شوند. نقطه مقابل جعبه سیاه شفافیت است، یعنی جستجو برای درک مستقیم مکانیزمی که یک مدل با آن کار می‌کند.

از آنجایی که مدل‌های یادگیری ماشین جعبه سیاه به طور فزاینده‌ای برای انجام پیش‌بینی‌های مهم در زمینه‌های حیاتی مورد استفاده قرار می‌گیرند، تقاضا برای شفافیت از سوی سهامداران مختلف در هوش مصنوعی افزایش می‌یابد . خطر در ایجاد و استفاده از تصمیماتی است که توجیه پذیر، مشروع نیست، یا به سادگی اجازه نمی دهد تا توضیحات دقیقی از رفتار آنها به دست آوریم . توضیحاتی که از خروجی یک مدل پشتیبانی می کند بسیار مهم است، به عنوان مثال، در پزشکی دقیق، جایی که متخصصان به اطلاعات بسیار بیشتری از مدل نیاز دارند تا یک پیش بینی باینری ساده برای پشتیبانی از تشخیص خود . نمونه های دیگر شامل وسایل نقلیه خودران در حمل و نقل، امنیت و امور مالی و غیره است.

به طور کلی، با توجه به تقاضـــای فزاینده برای هوش مصـــنوعی اخلاقی، انســـان‌ها نســـبت به اتخاذ تکنیک‌هایی که مستقیماً قابل تفسیر، قابل حمل و قابل اعتماد نیستند خودداری می‌کنند. معمولاً تصور می‌شـود که با تمرکز صـرف بر عملکرد، سـیسـتم ها به طور فزاینده ای مات می‌شـوند. این به این معنا درست است که بین عملکرد یک مدل و شفافیت آن یک معامله وجود دارد. با این حال، بهبود درک یک سـیسـتم می تواند منجر به اصـلاح کاسـتی های آن شـود. هنگام توسـعه یک مدل ML، در نظر گرفتن تفسیرپذیری به عنوان یک محرک طراحی اضافی می‌تواند به ۳ دلیل پیاده‌سازی آن را بهبود بخشد:

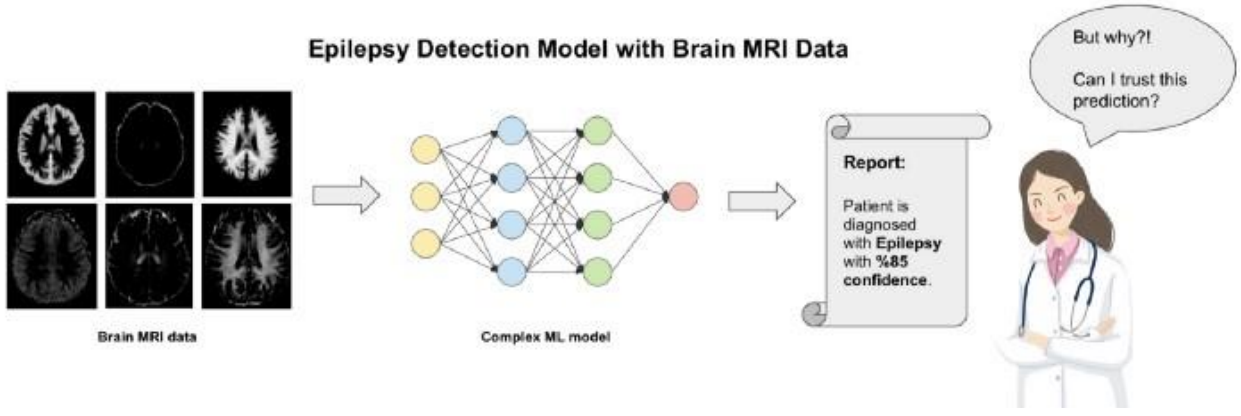
۱- تفسیرپذیری به اطمینان از بی طرفی در تصمیم گیری کمک می کند، یعنی شناسایی، و در نتیجه، تصحیح از سوگیری در مجموعه داده آموزشی.

۲- تفسیرپذیری با برجسته کردن اغتشاشات احتمالی متخاصم که می‌تواند پیش بینی را تغییر دهد باعث افزایش robustness مدل می‌شود.

۳- تفســـیرپذیری می‌تواند به‌عنوان بیمه‌ای عمل کند که تنها متغیرهای معنادار خروجی را استنتاج می‌کند، یعنی تضمین می‌کند که یک علیت واقعی در استدلال مدل وجود دارد.

۱-۲- اهمیت موضوع

برای اهمیت این موضوع به مثال زیر توجه کنید:



شکل ۱ – مثالی از XAI

فرض کنید پزشکی برای اینکه بفهمد مراجعه کننده آن دارای بیماری خاصی است یا خیر باید تصاویر مغزی گرفته شده از مراجعه کننده را بررسی کند.

برای این کار ما یک مدل پیچیده ماشین لرنینگ(شبکه عصبی عمیق) را ایجاد کرده ایم که با گرفتن عکس مغزی گرفته شده به عنوان ورودی شبکه عصبی به عنوان خروجی تشخیص میدهد که آیا این عکس نشان دهنده وجود بیماری در شخص است یا خیر.

این مدل که در تست‌ها بسیار موفق بوده است به طوری که صحت و دقت بالایی داشته را پزشک میخواهد حالا برای موارد واقعی که جان بیماری در خطر است و تشخیص پزشک در این مسئله بسیار حیاتی و مهم می باشد استفاده کند. پزشک عکس را به عنوان ورودی به مدل می‌دهد و مدل نتیجه ای را بیان میکند اما مشکل اینجاست که مدل به پزشک نمیگوید که چرا این نتیجه را گرفته است. اما پزشک چگونه می‌تواند در چنین تشخیص و تصمیم مهمی به نتیجه ای اعتماد کند که حتی نمیداند چگونه و چرا این تصمیم گرفته شده است.

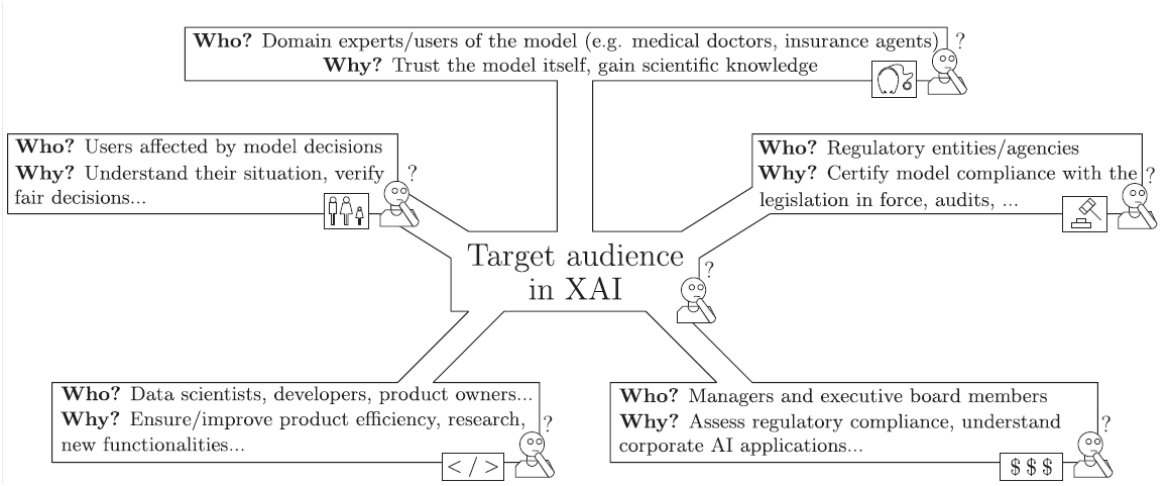
پس پزشک نیاز دارد که بداند چرا این تصمیم گرفته شده و مدل چگونه به این نتیجه رسیده است و اگر اینگونه نباشد پزشک نمیتواند به این مدل اعتماد کند و مدل نمیتواند عملیاتی شود. این مثال به واضحی بیان می‌دارد که چرا ما به تفسیر پذیری مدل نیاز داریم و اهمیت این موضوع تا چه حد می‌تواند باشد.

فرض کنید که سازمان های مهمی مانند سازمان های پزشکی ، نظامی ، قانونی و ... که تصمیمات بسیار مهم و حیاتی که جان هزاران انسان را میتواند به مخاطره بیاندازد و از مدل های هوش مصنوعی برای این تصمیم گیری ها میخواهند استفاده کنند چگونه میتوانند به مدل هایی که نمیتوانند تصمیمات آن ها را تفسیر کنند و بفهمند استفاده کنند.

بنابراین تقاضا برای هوش مصنوعی تفسیر پذیر یا XAI از سوی این سازمان ها هر روزه بیشتر می‌شود.

۱-۳- مخاطب و هدف توضیح

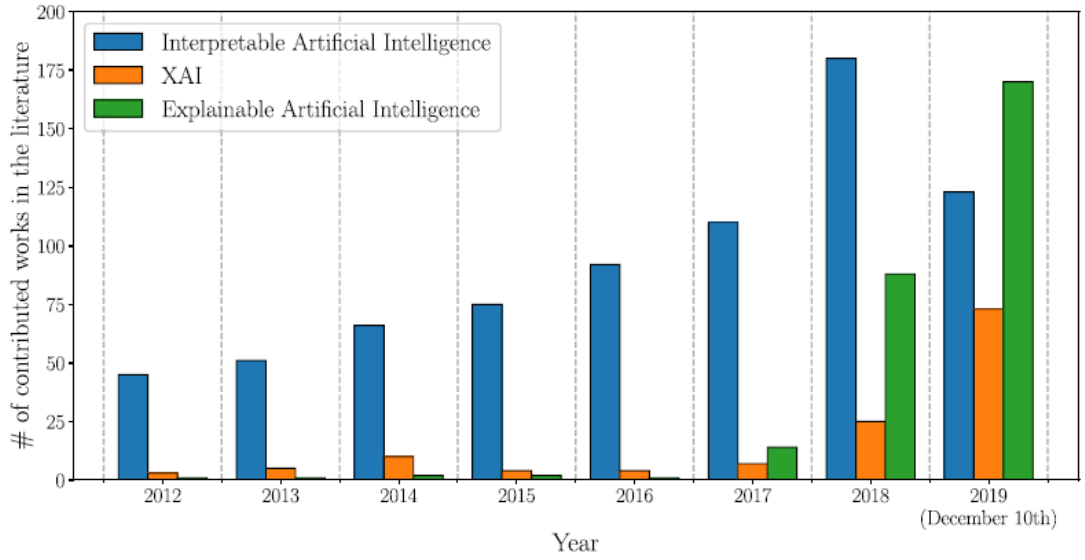
بسته به حوزه‌ی مورد استفاده هوش مصنوعی توضیح‌پذیری می‌تواند اهداف و مخاطب‌های گوناگونی داشته باشد. توضیحی که برای یک پزشک داده می‌شود با توضیحی که برای یک دانشمند داده یا مدیر شرکت گفته می‌شود متفاوت است.



شکل ۲ – نمودار مخاطب توضیح‌پذیری و چرایی

۱-۴- محبوبیت مشارکت XAI

همان‌طور که از شکل زیر مشخص است تحقیقاتی که با کلیدواژه‌های تفسیرپذیری و توضیح‌پذیری هوش مصنوعی در سال‌های اخیر انجام شده سیر صعودی داشته است و این موضوع نشان دهنده جذابیت و محبوبیت این حوزه می‌باشد.



شکل ۳ – نمودار افزایش مشارکت با کلیدواژه‌های مرتبط با XAI از سال ۲۰۱۲ تا ۲۰۱۹

فصل دوم

توضیح اصطلاحات و لغات تخصصی

۲-۱- مقدمه

در این بخش تفاوت ها و شباهت‌های بین اصطلاحاتی را که اغلب در XAI استفاده می‌شوند، توضیح می دهیم.

- قابل درک بودن (یا معادل آن، قابل فهم بودن): مشخصه یک مدل را نشان می دهد تا انسان عملکرد آن را بفهمد که مدل چگونه کار می کند بدون نیاز به توضیح ساختار داخلی آن یا ابزار الگوریتمی که مدل به وسیله آن داده ها را در داخل پردازش می کند.

- درک پذیری: زمانی که برای مدل های ML تصور می‌شود، قابل درک به تولنایی یک الگوریتم یادگیری برای نشان دادن دانش آموخته شده خود به شیوه ای قابل فهم برای انسان اشاره دارد.

- تفسیرپذیری: به عنوان توانایی توضیح دادن یا ارائه معنی در شرایط قابل درک برای انسان تعریف می‌شود.

- توضیح‌پذیری: توضیح‌پذیری با مفهوم تبیین به‌عنوان رابط بین انسان‌ها و تصمیم‌گیرنده مرتبط است که در عین حال، هم نماینده دقیق تصمیم‌گیرنده است و هم برای انسان قابل درک است.

۲-۲- شفافیت

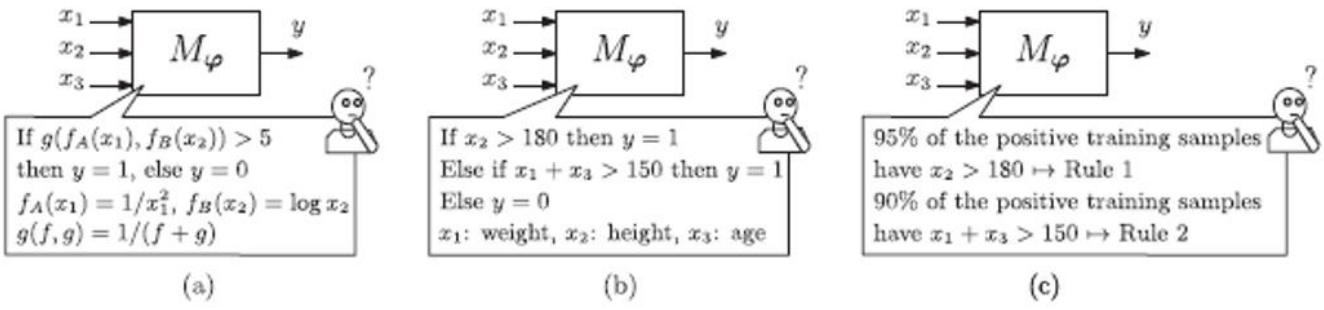
شفافیت: یک مدل در صورتی شفاف تلقی می‌شود که به خودی خود قابل درک باشد. از آنجایی که یک مدل می تواند درجات مختلفی از درک را داشته باشد.

مدل های شفاف به سه دسته تقسیم می شوند:

۱. مدل‌های شبیه‌سازی‌شده: نشان‌دهنده توانایی یک مدل شبیه‌سازی یا تفکر دقیق توسط یک انسان است، از این رو پیچیدگی در این کلاس جایگاه غالب دارد.

۲. مدل‌های تجزیه پذیر: مخفف توانایی توضیح هر یک از بخش های یک مدل (ورودی، پارامتر و محاسبه) است.

۳. شفافیت الگوریتمی: به روش های مختلفی قابل مشاهده است. این به توانایی کاربر برای درک فرآیند دنبال شده توسط مدل برای تولید هر خروجی داده شده از داده های ورودی آن می پردازد.



شکل ۴ – سطوح شفافیت

۲-۳- تعاریف

XAI(تعریف جدید): با توجه به مخاطبان، هوش مصنوعی قابل توضیح هوش مصنوعی است که جزئیات یا دلایلی را برای شفاف کردن یا درک آسان عملکرد خود ایجاد می کند.

ما اکنون تعاریفی را برای این اهداف XAI ترکیب و بر می‌شـــمریم تا اولین معیار طبقه‌بندی را تعیین کنیم:

- قابل اعتماد بودن: قابل اعتماد بودن را می توان به عنوان اطمینان از اینکه آیا یک مدل در مواجهه با یک مشـــکل معین همانطور که در نظر گرفته شـــده عمل می کند یا خیر تلقی می‌شود.

- علیت: یکی دیگر از اهداف مشـــترک برای توضـــیح پذیری، یافتن علیت در میان متغیرهای داده است.

- قابلیت انتقال: مدل ها همیشـــه با محدودیت هایی محدود می شـــوند که باید قابلیت انتقال یکپارچه آنها را فراهم کند.

- اطلاعات آموزی: مدل های ML قابل توضـــیح باید اطلاعاتی در مورد مشـــکلی که با آن برخورد می‌شود ارائه دهد.

- اعتماد: به عنوان تعمیم استحکام و پایداری، اعتماد همیشه باید بر اساس مدلی ارزیابی شود که در آن قابلیت اطمینان مورد انتظار است.

- انصاف: تجسم واضحی از روابط مؤثر بر نتیجه را نشان می دهد که امکان تحلیل عادلانه یا اخلاقی مدل در دست را فراهم می کند.

- دسترسی: به عنوان ویژگی که به کاربران نهایی اجازه می دهد تا بیشتر در فرآیند بهبود و توسعه یک مدل ML خاص مشارکت کنند.

- تعامل: شامل تولنایی یک مدل برای تعامل با کاربر به عنوان یکی از اهداف مورد نظر یک مدل ML قابل توضیح است.

- آگاهی از حریم خصوصی: یکی دیگر از اهداف مشترک برای توضیح پذیری، یافتن علیت بین متغیرهای داده است.

XAI Goal	Main target audience (Fig. 2)
Trustworthiness	Domain experts, users of the model affected by decisions
Causality	Domain experts, managers and executive board members, regulatory entities/agencies
Transferability	Domain experts, data scientists
Informativeness	All
Confidence	Domain experts, developers, managers, regulatory entities/agencies
Fairness	Users affected by model decisions, regulatory entities/agencies
Accessibility	Product owners, managers, users affected by model decisions
Interactivity	Domain experts, users affected by model decisions
Privacy awareness	Users affected by model decisions, regulatory entities/agencies

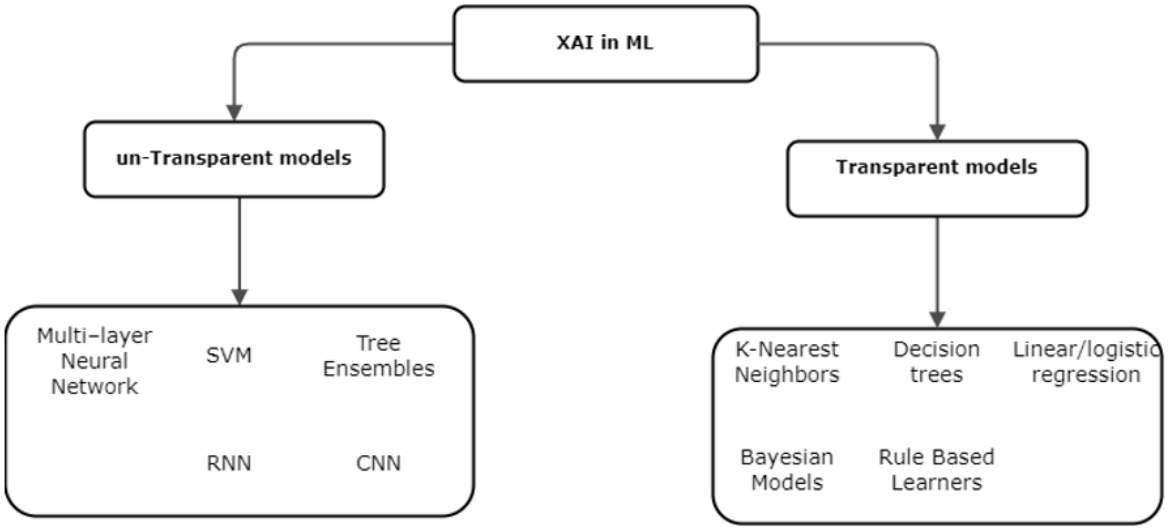
شکل ۵ – رابطه‌ی اهداف و مخاطب هدف

فصل سوم

XAI در مدل‌های یادگیری ماشین

۳-۱- مقدمه

همانطور که در شکل زیر دیده می‌شود، بر اساس پارامتر توضیح پذیری مدل های یادگیری ماشین به دو دسته شفاف و غیر شفاف تقسیم می شوند.



شکل ۶ – تقسیم‌بندی مدل‌ها از نظر شفافیت

دسته‌ی شفاف متشکل است از مدل هایی مانند:

- رگرسیون خطی/ لجستیک[1]
- درخت تصمیم گیری[2]
- K نزدیکترین همسایه[3]

[1] Linear/logistic regression

[2] Decision tree

[3] K-nearest neighbors

- مدل های مبتنی بر قانون[1]
- مدل های بیز[2]

همچنین دسته غیر شفاف شامل مدل های جعبه سیاه مانند مدل های زیر می باشد:

- ماشین بردار پشتیبانی[3]
- شبکه های عصبی چند لایه ای/ بازگشتی/ پیچشی[4]
- درخت های جمعی[5]

در واقع در این بخش به دنبال این هستیم که مدل های شفاف معروف را بر اساس سه معیار قابل تصور بودن، تجزیه پذیری و شفافیت الگوریتمی بررسی کنیم و همچنین روش هایی برای افزایش توضیح پذیری در مدل های غیر شفاف را ارائه دهیم.

۳-۲- رگرسیون خطی/ لجستیک

این مدل قابل تصور می باشد، زیرا در واقع خطی یا ابر صفحه ای در فضا است و مشخص است که این خط به این خاطر انتخاب شده است که میانگین فاصله آن با نقاط دیگر به صورت کمینه می باشد. همچنین از نظر تجزیه پذیری، در صورت اینکه از مهندسی ویژگی استفاده نکرده باشیم، ویژگی ها یا ورودی های ما با معنی و مشخص خواهند بود و همچنین پارامتر های مدل ما واضح هستند و محاسبات مدل از ورودی ها و پارامتر ها جدا است و با آن ها در هم نیامیخته است. در نهایت این مدل دارای شفافیت الگوریتمی کمی می باشد، زیرا پایه آن محاسبات ریاضی نسبتا پیچیده ای می باشد و به راحتی نمی توان تصور کرد که خروجی توسط این مدل به چه نحوی تولید شده است.

[1] Rule based models

[2] Bayesian models

[3] Support vector machine

[4] Multi-layer/Recurrent/Convolutional neural networks

[5] Tree ensembles

۳-۳- درخت تصمیم گیری

درخت تصمیم گیری را به راحتی می توان تصور کرد، زیرا این مدل در واقع شامل مجموعه ای از شروط می باشد که با توجه به آن ها خروجی مدل تولید می شود که به راحتی می توان آن را به وسیله یک درخت متصور کرد. همچنین این مدل دارای تجزیه پذیری بالایی می باشد، زیرا ورودی و پارامتر های آن کاملا مشخص هستند و محاسبات آن تغییراتی روی داده ها و ورودی ها اصلا ایجاد نمی کند و دارای خوانایی[1] بالایی می باشد. در کل می توان قدم هایی که طی می شود تا مدل خروجی تولید کند را به راحتی تصور کرد و در واقع این مدل دارای شفافیت الگوریتمی بالایی می باشد.

۳-۴- K نزدیکترین همسایه

این مدل بسیار قابل تصور می باشد، زیرا متشکل از خطوط یا ابر صفحه هایی در فضا است که داده ها را به دسته هایی تقسیم کرده است و شکلی محدب تولید کرده است. همچنین دارای تجزیه پذیری بالایی می باشد، چرا که متغیر ها از تابع شباهت و خروجی مدل کاملا جدا هستند و اجزای مختلف مدل از هم به راحتی جدا پذیر هستند. اما این مدل دارای شفافیت الگوریتمی کمی می باشد چرا که ممکن است تابع شباهت یا معیاری که برای شباهت استفاده می شود پیچیده باشد و همچنین تعداد متغیر ها و ورودی ها زیاد می باشد و به خاطر همین مقایسه ها و محاسبه شباهت ها به راحتی در ذهن قابل تصور نمی باشد و به راحتی نمی توان قدم هایی که طی می شود تا مدل خروجی تولید شود را به راحتی در نظر گرفت.

۳-۵- مدل های مبتنی بر قانون

این مدل ها با توجه به اینکه تعداد قانون ها زیاد باشد و در واقع قانون ها به صورت اتوماتیک تولید شده باشد یا تعداد قانون ها کم باشد و این قانون ها بر اساس منطق باشند، با توجه به سه معیار معرفی شده

[1] Readability

ممکن اســت دارای شــفافیت کم یا بالایی باشــند و در واقع شــفافیت این مدل با تعداد قوانین آن رابطه عکس و با منطقی بودن قوانین آن رابطه مستقیم دارد.

۳-۶- مدل های بیز

شفافیت این مدل ها توجه به این سه معیار معرفی شده با پیچیدگی این مدل ها رابطه عکس دارد و در واقع هر چه گراف رابطه متغیر های تصادفی پیچیده تر و تعداد آن ها بیشــتر باشــد، شــفافیت این مدل کمتر می باشد.

Model	Transparent ML Models			Post-hoc analysis
	Simulatability	Decomposability	Algorithmic Transparency	
Linear/Logistic Regression	Predictors are human readable and interactions among them are kept to a minimum	Variables are still readable, but the number of interactions and predictors involved in them have grown to force decomposition	Variables and interactions are too complex to be analyzed without mathematical tools	Not needed
Decision Trees	A human can simulate and obtain the prediction of a decision tree on his/her own, without requiring any mathematical background	The model comprises rules that do not alter data whatsoever, and preserves their readability	Human-readable rules that explain the knowledge learned from data and allows for a direct understanding of the prediction process	Not needed
K-Nearest Neighbors	The complexity of the model (number of variables, their understandability and the similarity measure under use) matches human naive capabilities for simulation	The amount of variables is too high and/or the similarity measure is too complex to be able to simulate the model completely, but the similarity measure and the set of variables can be decomposed and analyzed separately	The similarity measure cannot be decomposed and/or the number of variables is so high that the user has to rely on mathematical and statistical tools to analyze the model	Not needed
Rule Based Learners	Variables included in rules are readable, and the size of the rule set is manageable by a human user without external help	The size of the rule set becomes too large to be analyzed without decomposing it into small rule chunks	Rules have become so complicated (and the rule set size has grown so much) that mathematical tools are needed for inspecting the model behaviour	Not needed
Bayesian Models	Statistical relationships modeled among variables and the variables themselves should be directly understandable by the target audience	Statistical relationships involve so many variables that they must be decomposed in marginals so as to ease their analysis	Statistical relationships cannot be interpreted even if already decomposed, and predictors are so complex that model can be only analyzed with mathematical tools	Not needed
Tree Ensembles	✗	✗	✗	Needed: Usually *Model simplification* or *Feature relevance* techniques
Support Vector Machines	✗	✗	✗	Needed: Usually *Model simplification* or *Local explanations* techniques
Multi-layer Neural Network	✗	✗	✗	Needed: Usually *Model simplification*, *Feature relevance* or *Visualization* techniques
Convolutional Neural Network	✗	✗	✗	Needed: Usually *Feature relevance* or *Visualization* techniques
Recurrent Neural Network	✗	✗	✗	Needed: Usually *Feature relevance* techniques

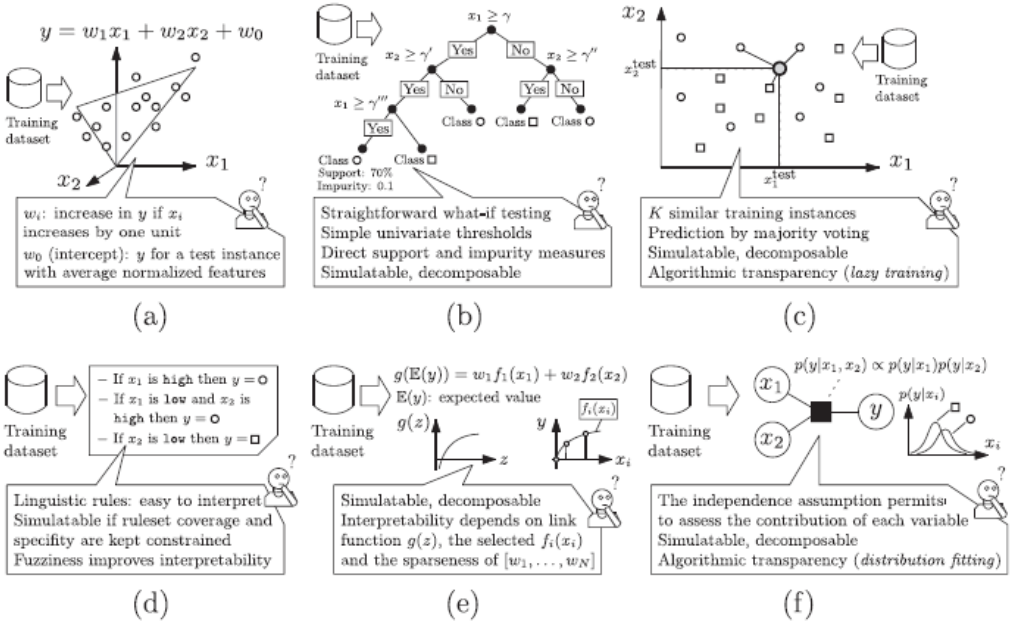
شکل ۷ – تقسیم‌بندی مدل‌های یادگیری ماشین و سطوح شفافیت

۳-۷- روش‌های افزایش توضیح‌پذیری

برای افزایش توضـــیح پذیری در مدل های غیر شـــفاف از روش های تعقیبی اســـتفاده می کنیم که این روش ها شامل اند از:

- توضیح نوشتاری
- توضیح تصویری
- توضیح محلی
- توضیح با مثال
- توضیح با ساده سازی
- توضیح با بررسی ارتباط ویژگی ها با خروجی

این روش ها به ترتیب در شکل زیر دیده می شوند.



شکل ۸ – روش‌های توضیح‌پذیری

- توضیح نوشتاری

در این روش با ارائه توضـیحات نوشـتاری در مورد اینکه چگونه خروجی توسـط مدل تولید می شود، توضیح پذیری مدل را افزایش می دهیم، که خیلی روش مناسبی نمی باشد.

- توضیح تصویری

با ارائه تصـوری از رفتار مدل، توضـیح پذیری آن را افزایش می دهیم. برای مثال با می توانیم با استفاده از یک درخت تصمیم رفتار مدل را نشان دهیم، به طوری که خروجی درخت تصمیم با خروجی مدل مرتبط باشد.

- توضیح محلی

با تقسیم فضای جواب به قسمت هایی کوچک تر و توضیحات در مورد رفتار مدل در این قسمت ها، می توانیم دید بهتری از مدل داشـته باشـیم. برای مثال می توانیم مدل را در هر یک از این قسمت ها استفاده کنیم و حال به وسیله روش های جمعی یک خروجی کلی داشته باشیم.

- توضیح با مثال

می توانیم به با استفاده از ورودی هایی به عنوان مثال و خروجی مدل با استفاده از آن ها، رفتار مدل را توضیح دهیم. برای مثال با استفاده از داده های آموزشی حیاتی و بردار های پشتیبانی، رفتار مدل های ماشین بردار پشتیبانی را مشخص کنیم.

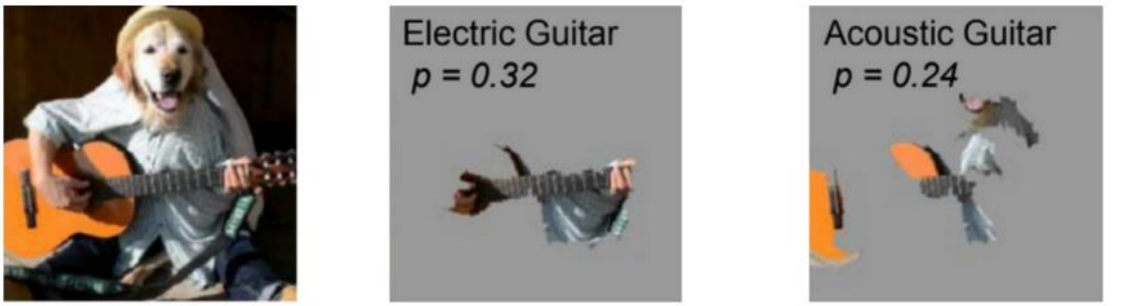
- توضیح با ساده سازی

با ساده سازی مدل تا جایی که دقت خیلی فدا نشود، می توانیم توضیح پذیری مدل را افزایش بدهیم و در واقع این طبیعت مدل های یادگیری ماشـین می باشـد. برای مثال در شـبکه های عصبی با حذف بعضی از ارتباطات و ساده سازی مدل می توانیم توضیح پذیری آن را افزایش دهیم.

- توضیح با بررسی ارتباط ویژگی ها با خروجی

با بررسی ارتباط ویژگی ها با خروجی و اینکه کدام ورودی ها بیشتر در خروجی تاثیر دارند، می توانیم رفتار مدل را بهتر نشان دهیم و توضیح پذیری آن را افزایش بدهیم.

برای مثال همان‌طور که در شکل زیر دیده می‌شود با استفاده از چهارچوب LIME[1] می توانیم نشان دهیم که چه بخش‌هایی از تصویر در دسته‌بندی تصویر بیشتر تاثیر دارند.

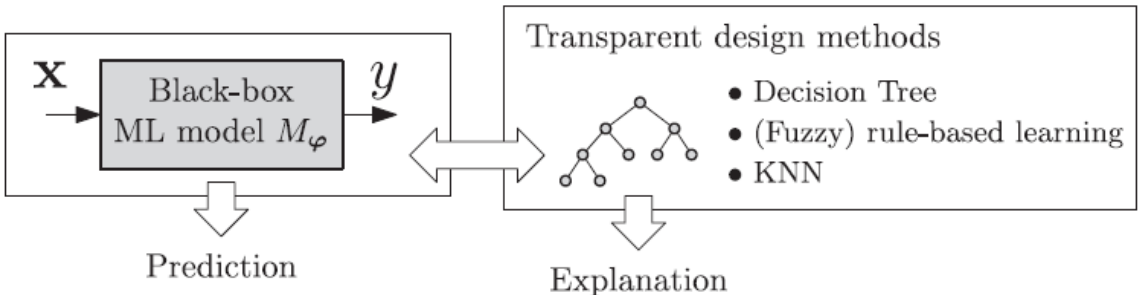


شکل ۹ – استفاده از چارچوب LIME

۳-۸- مدل های ترکیبی شفاف و جعبه سیاه

با استفاده از روش های معرفی شده برای افزایش توضیح پذیری مدل های غیر شفاف، می توانیم مدل شفافی معادل با مدل غیر شفاف یا جعبه سیاه بسازیم که رفتاری و دقتی شبیه به آن دارد و دارای توضیح پذیری بیشتری می باشد.

برای مثال همانطور که در شکل زیر دیده می‌شود می توانیم درخت تصمیمی معادل با شبکه عصبی بسازیم و توضیح پذیری مدل غیر شفاف را با این روش افزایش دهیم.



شکل ۱۰ – نمودار مدل ترکیبی

[1] Local Interpretable Model-Agnostic Explanations

فصل چهارم

فرصت‌های تحقیق و هوش مصنوعی مسئول

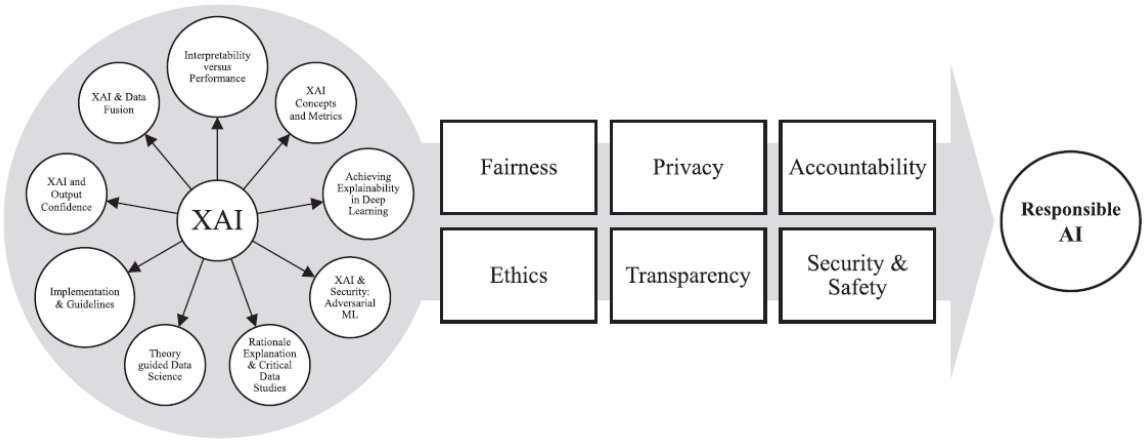
۴-۱- فرصت های تحقیق در XAI

در حوزه XAI فرصت های زیادی برای تحقیق موجود می باشد. یکی از این فرصت ها تحقیق در روش هایی برای رسیدن به تعادل مناسبی بین توضیح پذیری و عملکرد یک مدل می باشد. دیگر فرصت ها شامل از تحقیقات در مورد معنا و شرح دقیق تری از توضیح پذیری در هوش مصنوعی می باشند.

۴-۲- هوش مصنوعی مسئول[1]

به‌طور خلاصه اگر ۶ مورد انصاف، حریم شخصی، پاسخ‌گویی، اخلاق، شفافیت و ایمنی و امنیت در یک سیستم مبتنی بر هوش مصنوعی برقرار باشد به مفهوم هوش مصنوعی مسئول دست می‌یابیم.

بعد از حل چالش‌های ۶ مورد مطرح شده، مفهوم هوش مصنوعی مسئول به پیشرفت و کارگیری از هوش مصنوعی کمک کرده و می‌تواند به مشتریان این صنعت اطمینان لازم را برای استفاده بدهد



شکل ۱۱ – نمودار هوش مصنوعی مسئول

[1] Responsible AI

فصل پنجم

چرا باید به تو اعتماد کنم؟

مدل توضیح پذیر LIME

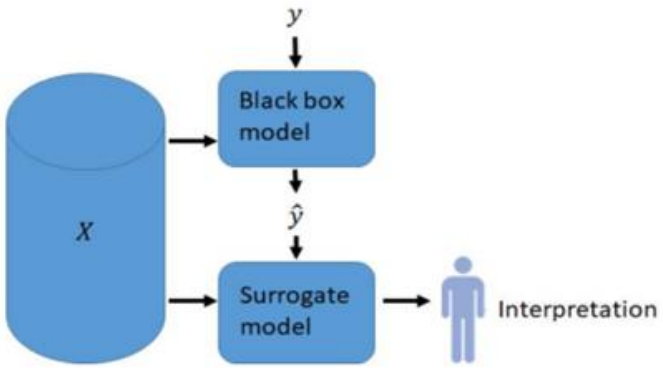
مدل‌های جایگزین محلی مدل‌های قابل تفسیری هستند که برای توضیح پیش‌بینی‌های مدل‌های جعبه سیاه یادگیری ماشین استفاده می‌شوند. «توضیحات مستقل از مدل تفسیرپذیر محلی» [۱۲] به‌اختصار LIME روشی است که در آن پیاده‌سازی ملموسی از مدل‌های جایگزین محلی پیشنهاد می‌شود. مدل‌های جایگزین برای تقریب زدن پیش‌بینی‌های مدل جعبه سیاه آموزش می‌بینند و به‌جای آموزش یک مدل جانشین کلی، LIME بر آموزش مدل‌های جایگزین محلی برای توضیح پیش‌بینی‌ها تمرکز می‌کند.

ایده‌ی الگوریتم LIME کاملاً شهودی است. به این صورت که به داده‌های آموزشی توجه نمی‌شود و فقط مدل جعبه سیاهی را گرفته که می‌توان نقاط داده را وارد کرد و پیش‌بینی‌های مدل را به دست آورد و جعبه را بررسی کرد. هدف از این کار درک این موضوع است که چرا مدل یادگیری ماشین چنین پیش‌بینی‌ای انجام داده است. LIME آزمایش می‌کند که وقتی تغییراتی از داده‌های خود را در مدل یادگیری ماشین می‌دهید، چه اتفاقی برای پیش‌بینی‌ها می‌افتد؛ بنابراین LIME یک مجموعه‌داده جدید از پیش‌بینی مدل جعبه سیاه تولید می‌کند که شامل نمونه‌هایی

است. در این مجموعه‌داده جدید LIME سپس یک مدل قابل تفسیر را آموزش می‌دهد که با نزدیکی نمونه‌های نمونه‌برداری شده به نمونه موردنظر وزن‌دهی می‌شود. مدل قابل تفسیر می‌تواند هر چیزی از جنس مدل‌های شفاف برای مثال درخت تصمیم باشد. مدل آموخته شده باید تقریب خوبی از پیش‌بینی‌های مدل یادگیری ماشین به‌صورت محلی باشد، اما لزومی ندارد که یک تقریب کلی خوب باشد. به این نوع دقت، وفاداری محلی نیز می‌گویند.

دستور العمل برای آموزش مدل های جایگزین محلی به شرح زیر است:

۱) نمونه مورد علاقه خود را که می خواهید توضیحی درباره پیش بینی جعبه سیاه آن داشته باشید را انتخاب کنید.
۲) مجموعه داده خود را آشفته کنید و پیش بینی های جعبه سیاه را برای این نقاط جدید دریافت کنید.
۳) نمونه های جدید را با توجه به نزدیکی آنها به نمونه مورد نظر وزن کنید.
۴) یک مدل وزن دار و قابل تفسیر روی مجموعه داده با تغییرات آموزش دهید.
۵) پیش بینی را با تفسیر مدل محلی توضیح دهید.



شکل ۱۲- شمای کلی از مدل‌های توضیح پذیر

فصل ششم

نتیجه‌گیری و جمع بندی

به عنوان جمع‌بندی، ابتدا مقدمه‌ای از هوش مصـنوعی توضـیح‌پذیر و اهمیت این حوزه و مخاطب و هدف آن ارائه گشت و سپس به مفاهیم و تعاریف اولیه و اصطلاحات مورد نیاز برای علمی‌سازی XAI و تعریف شفافیت و سطوح آن پرداخته شد. سپس به بررسی توضیح‌پذیری مدل‌های یادگیری ماشین و روش‌های مختلف توضـیح مبادرت نموده شـد. چالش‌های تحقیق در این حوزه‌ی جدید ارائه و مفهوم جدید هوش مصنوعی مسئول معرفی شد و در آخر مدل توضیح پذیر محلی Lime مورد آشنایی قرار گرفت.

به‌عنوان نتیجه‌گیری، معرفی این حوزه‌ی جدید و عملی‌سـازی برای کارهای تحقیقاتی آینده و آشـنایی با چالش‌های این حوزه در راستای کمک به صنعت هوش مصنوعی است. به طوری که مشتریان سیستم‌های هوش مصنوعی اطمینان مورد نیاز برای خرید و استفاده از هوش مصنوعی را حاصل کنند. هوش مصنوعی توضیح پذیر سعی بر حل این مشکلات و چالش‌ها دارد.

فصل هفتم

کد و تفسیر آن به صورت گام به گام

Dataset

هر رکورد در پایگاه داده یک حومه یا شـهر بوسـتون را توصـیف می‌کند. داده‌ها از منطقه آماری شـهری استاندارد بوستون (SMSA) در سال ۱۹۷۰ استخراج شد. ویژگی‌ها به شرح زیر تعریف شده‌اند (برگرفته از مخزن یادگیری ماشین UCI1):

CRIM: نرخ جرم سرانه بر اساس شهر

ZN: نسبت زمین مسکونی پهنه‌بندی شده برای زمین‌های بیش از ۲۵۰۰۰ فوت مربع.

INDUS: نسبت هکتارهای تجاری غیر خرده‌فروشی در هر شهر

CHAS: متغیر سـاختگی رودخانه چارلز (= ۱ اگر مسـیر به رودخانه محدود می‌شـود؛ ۰ در غیر این صورت)

NOX: غلظت اکسیدهای نیتریک (قسمت در هر ۱۰ میلیون) مجموعه‌داده ۱۲۴ را بارگیری کنید

RM: میانگین تعداد اتاق در هر خانه

AGE: نسبت واحدهای تحت اشغال ساخته شده قبل از سال ۱۹۴۰

DIS: فواصل وزنی تا پنج مرکز استخدامی بوستون

RAD: شاخص دسترسی به بزرگراه‌های شعاعی

مالیات: نرخ مالیات بر دارایی باارزش کامل به‌ازای هر ۱۰۰۰۰ دلار

PTRATIO: نسـبت دانش‌آموز به معلم بر اسـاس شـهر ۱۲. B: 1000 (Bk-0.63)2 که در آن Bk نسبت سیاه‌پوستان بر اساس شهر است ۱۳. LSTAT: ٪ وضعیت پایین‌تر جمعیت

MEDV: میانگین ارزش خانه‌های تحت اشغال در ۱۰۰۰ دلار

می‌بینیم که ویژگی‌های ورودی ترکیبی از واحدها دارند.

همان‌طور که در شکل زیر مشاهده می‌گردد کتابخانه‌ها فراخوانی شده‌اند فیلدها و ویژگی ستون‌ها به فایل اضافه شده‌اند و با استفاده از دستوراتی عملیات پیش‌پردازش‌ها و عملیات مدیریت‌کردن داده‌های مفقوده صورت‌گرفته است تا داده‌ها آماده مدل زدن باشند.

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
```

```
[2] # Ignore warnings
    import warnings
    warnings.filterwarnings('ignore')
```

```
[3] #Lets load the dataset and sample some
    column_names = ['CRIM', 'ZN', 'INDUS', 'CHAS', 'NOX', 'RM', 'AGE', 'DIS', 'RAD', 'TAX', 'PTRATIO', 'B', 'LSTAT', 'MEDV']

    df = pd.read_csv('/content/housing.csv', header=None, delimiter=r"\s+", names=column_names)
    df.head()
```

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222.0	18.7	396.90	5.33	36.2

با استفاده از دستور زیر دیتا تایپ هرکدام از فیلدها مشخص شده است و همچنین مشخص شده است که فیلد خالی وجود ندارد.

```
[4] df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 506 entries, 0 to 505
Data columns (total 14 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   CRIM     506 non-null    float64
 1   ZN       506 non-null    float64
 2   INDUS    506 non-null    float64
 3   CHAS     506 non-null    int64
 4   NOX      506 non-null    float64
 5   RM       506 non-null    float64
 6   AGE      506 non-null    float64
 7   DIS      506 non-null    float64
 8   RAD      506 non-null    int64
 9   TAX      506 non-null    float64
 10  PTRATIO  506 non-null    float64
 11  B        506 non-null    float64
 12  LSTAT    506 non-null    float64
 13  MEDV     506 non-null    float64
dtypes: float64(12), int64(2)
memory usage: 55.5 KB
```

برای راحتی کار و تسریع مدل‌سازی ۶ تا از مهم‌ترین ویژگی‌ها را انتخاب شده است و تارگت مسئله نیز که همان میانگین ارزش خانه‌های تحت اشغال * ۱۰۰۰ دلار هست مشخص گردیده است.

فیلدها و ویژگی‌های انتخاب شده به فارسی در کد آمده است.

```
[6]

    # Declare feature vector and target variable
    X = df[['LSTAT','RM','NOX','PTRATIO','DIS','AGE']]
    y = df['MEDV']
```

در اینجا، من 6 متغیر زیر را به عنوان بردار ویژگی برای راحتی انتخاب کرده ام.

```
1 LSTAT - وضعیت پایین تر جمعیت
2 RM - میانگین تعداد اتاق در هر مسکن
3 NOX - غلظت اکسیدهای نیتریک (قسمت در هر 10 میلیون)
4 PTRATIO - نسبت دانش آموز به معلم بر اساس شهر
5 DIS - فاصله وزنی تا پنج مرکز استخدامی بوستون
6 AGE - 1940 نسبت واحدهای تحت اشغال ساخته شده قبل از سال
```

متغیر هدف (ام ای دی وی) است که مخفف میانگین ارزش خانه‌های تحت اشغال است

داده‌ها با اســـتفاده از کتابخانه sklearn.model_selection داده‌ها به قســـمت‌های Train و Test به‌صـورت رندوم تقسـیم می‌شـوند و ۷۰ درصـد داده‌ها را برای آموزش و ۳۰ درصـد برای تسـت تقسـیم می‌کنیم. سـپس با اسـتفاده از کتابخانه sklearn.ensemble مدل تجمعی جنگل تصـادفی روی دادهای آموزشی با استفاده از یک‌سری از پارامترهای مدنظر پیاده‌سازی می‌کنیم.

مدل آماده شـــده و خروجی‌های پیش‌بینی‌شـــده به دســـت می‌آید ســـپس با اســـتفاده از کتابخانه sklearn.metric شاخص mean_squared_error به دست می‌آید که از آن جذر گرفته

(به توان 0.5 رسانده) loss و test score مدل برابر ۴.۴۴ می‌شود.

```
[7]  # Split the data into train and test data:
     from sklearn.model_selection import train_test_split
     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = 0)
```

```
[8]  # Build the model with Random Forest Regressor :
     from sklearn.ensemble import RandomForestRegressor
     model = RandomForestRegressor(max_depth=6, random_state=0, n_estimators=10)
     model.fit(X_train, y_train)
```

```
     RandomForestRegressor(max_depth=6, n_estimators=10, random_state=0)
```

```
[9]  y_pred = model.predict(X_test)
```

```
[10] from sklearn.metrics import mean_squared_error
     mse = mean_squared_error(y_test, y_pred)**(0.5)
     mse
```

```
     4.438832852155771
```

در google colab کتابخانه lime به‌صـــورت پیش‌فرض وجود ندارد؛ بنابراین با دســـتور pip آن را به framework خود اضافه می‌کنیم و با استفاده از دستور import آنها را فراخوانی می‌کنیم.

باتوجه‌به جدولی بودن مجموعه‌داده ما در اینجا مدل lime_tabular را فراخوانی می‌کنیم.

```
[11] !pip install lime

     Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
     Requirement already satisfied: lime in /usr/local/lib/python3.8/dist-packages (0.2.0.1)
     Requirement already satisfied: numpy in /usr/local/lib/python3.8/dist-packages (from lime) (1.21.6)
     Requirement already satisfied: tqdm in /usr/local/lib/python3.8/dist-packages (from lime) (4.64.1)
     Requirement already satisfied: scipy in /usr/local/lib/python3.8/dist-packages (from lime) (1.7.3)
     Requirement already satisfied: scikit-learn>=0.18 in /usr/local/lib/python3.8/dist-packages (from l
     Requirement already satisfied: scikit-image>=0.12 in /usr/local/lib/python3.8/dist-packages (from l
     Requirement already satisfied: matplotlib in /usr/local/lib/python3.8/dist-packages (from lime) (3.
     Requirement already satisfied: networkx>=2.0 in /usr/local/lib/python3.8/dist-packages (from scikit
     Requirement already satisfied: pillow!=7.1.0,!=7.1.1,>=4.3.0 in /usr/local/lib/python3.8/dist-packa
     Requirement already satisfied: tifffile>=2019.7.26 in /usr/local/lib/python3.8/dist-packages (from
     Requirement already satisfied: imageio>=2.3.0 in /usr/local/lib/python3.8/dist-packages (from sciki
     Requirement already satisfied: PyWavelets>=1.1.1 in /usr/local/lib/python3.8/dist-packages (from sc
     Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.8/dist-packages (from ma
     Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.8/dist-packages (from matplot
     Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in /usr/local/lib/python3.8
     Requirement already satisfied: python-dateutil>=2.1 in /usr/local/lib/python3.8/dist-packages (from
     Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.8/dist-packages (from
     Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.8/dist-packages (from scikit-
     Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.8/dist-packages (from python-date
```

```
[12] import lime
     import lime.lime_tabular
```

با استفاده از تابع LimeTabularExplainer برای توضیح طبقه‌بندی‌کننده‌هایی که از داده‌های جدولی (ماتریس)هستند استفاده شده است. در این تابع از پارامترهای متفاوتی استفاده شده است که

به ترتیب مقادیر داده‌های آموزش شـــده ، آن ۶ ویژگی (که در بالا انتخاب کرده بودیم) متارگت و هدف مسئله ، verbose=1 که نشان‌دهنده خطی‌بودن مدل است و Mode که در اینجا رویکرد Regression مدنظر ماست انتخاب شده‌اند و مدل توضیح‌پذیر Lime آموزش‌دیده است.

ســـپس یک نمونه (به عنوان مثال نمونه ۵ ام) انتخاب شـــده و به مدل Lime که در بالا آموزش دادیم داده شده است که با استفاده از آن قیمت پیش‌بینی شده محلی به‌دست‌آمده است.

class **lime.lime_tabular.LimeTabularExplainer**(*training_data, mode='classification', training_labels=None, feature_names=None, categorical_features=None, categorical_names=None, kernel_width=None, kernel=None, verbose=False, class_names=None, feature_selection='auto', discretize_continuous=True, discretizer='quartile', sample_around_instance=False, random_state=None, training_data_stats=None*)

```
[13] # LIME has one explainer for all the models
     explainer = lime.lime_tabular.LimeTabularExplainer(X_train.values, feature_names=X_train.columns.values.tolist(),
                                                        class_names=['MEDV'], verbose=True, mode='regression')
```

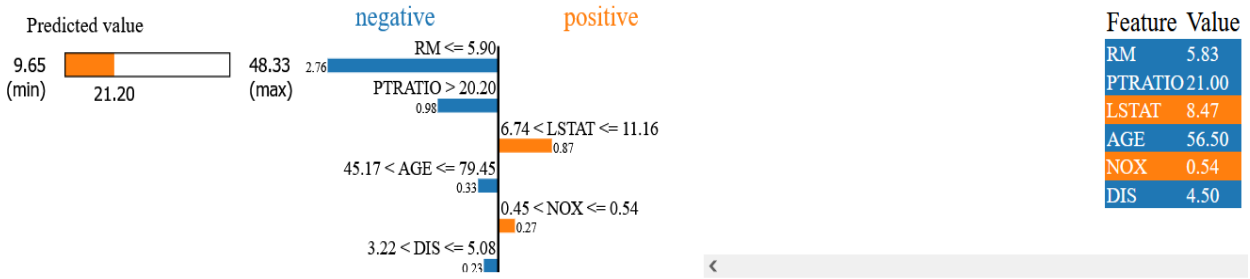
در اینجا، من نمونه 5ام دیتاست را انتخاب می کنم و از آن برای توضیح پیش بینی ها استفاده می کنم و وزن و علت هر یک از ویژگی‌ها برای این قیمت پیش‌بینی شده مشخص می‌گردد

```
[14] # Choose the 5th instance and use it to predict the results
     j = 5
     exp = explainer.explain_instance(X_test.values[j], model.predict, num_features=6)

     Intercept 24.768911172989316
     Prediction_local [21.6001998]
     Right: 21.200831768172357
```

سپس مدل را با استفاده از تابع show_in_notebook و در mode که نمودارها را نیز نشان دهد قرار داده‌ایم. حال با مشخص‌شدن وزن هر یک از ویژگی‌ها و تأثیرات مثبت یا منفی هریک از آنها در قیمت نهایی منزل انتخاب شده به دست می‌آید تا دلایل و تفسیر قیمت منزل پیش‌بینی‌شده به دست بیاید.

```
[15] # Show the predictions
     exp.show_in_notebook(show_table=True)
```




```
[16] exp.as_list()
```

```
[('RM <= 5.90', -2.7638041433417593),
 ('PTRATIO > 20.20', -0.981049149580379),
 ('6.74 < LSTAT <= 11.16', 0.8677483690962801),
 ('45.17 < AGE <= 79.45', -0.32685120309898263),
 ('0.45 < NOX <= 0.54', 0.269501645444881),
 ('3.22 < DIS <= 5.08', -0.2342568959153656)]
```

تفسیر

ارزش پیش بینی شده قیمت مسکن 21.48 است.

تأثیر منفی بر قیمت پیش‌بینی‌شده مسکن دارند RM، DIS، AGE و PTRATIO تأثیر مثبت دارند در حالی که NOX و LSTAT متغیرهای

همه ارزش ها به هزار دلار است

منابع و مراجع

1) Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. and Chatila, R., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, pp.82-115.

2) Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).