

فهم مسئله :

(۱) انگیزه اصلی چنین پروژه هایی چیست؟

انگیزه اصلی پروژه هایی که در آنها از الگوریتم یادگیری ماشینی مانند رگرسیون خطی استفاده می شود، می تواند متفاوت باشد اما به طور کلی چند هدف مهم می توانند عبارت باشند از:

۱- **پیش بینی و تحلیل :** استفاده از داده های تاریخی برای پیش بینی رفتارها یا ارزش های آینده به عنوان مثال پیش بینی قیمت خانه ها بر اساس ویژگی های مختلف می تواند به سرمایه گذاران املاک و مستغلات، خریداران خانه و سیاستگذاران کمک کند تا تصمیمات آگاهانه تری بگیرند. (به خریداران کمک می کند تا قیمت منطقی را برای خرید خانه تخمین بزنند. و به فروشندگان کمک می کند تا قیمت مناسبی برای فروش خانه تعیین کنند).

۲- **تصمیم گیری مبتنی بر داده :** در دنیای امروز که داده ها به یک منبع ارزشمند بدیل شده اند، توانایی تجزیه و تحلیل داده ها می تواند به سازمان ها کمک کند تا تصمیم های بهتری بگیرند. به عنوان مثال تصمیم گیری در مورد محل ساخت خانه های جدید بر اساس تقاضای پیش بینی شده.

۳- **بهینه سازی منابع:** با پیش بینی دقیق تر، سازمان ها و افراد می توانند منابع خود را به شیوه ای کارآمدتر تخصیص دهند. به عنوان مثال توسعه دهندگان املاک می توانند تصمیم بگیرند که در کدام مناطق سرمایه گذاری کنند که بازده بیشتری داشته باشد.

۴- **شخصی سازی و بهبود خدمات:** در برخی موارد الگوریتم ها می توانند به شخصی سازی خدمات برای کاربران نهایی کمک کنند. مانند پیشنهاد دادن خانه هایی که با توجه به ترجیحات و بودجه آن ها مناسب است.

۵- **کشف الگوها و روندها :** تحلیل داده ها می تواند به شناسایی الگوها و روندهایی که قبلاً ناشناخته بودند کمک کند، ممکن است این اطلاعا برای توسعه سیاست های جدید و یا غیر استراتژی فعلی مفید باشد.

۶- **تحلیل بازار مسکن:** درک بهتر عوامل موثر بر قیمت خانه در یک منطقه خاص، شناسایی روندها و الگوهای قیمت در طول زمان. این پروژه ها به طور کلی به دنبال افزایش دقت و کارایی در فرایندهای تجاری، علمی و اداری هستند و می توانند تأثیر قابل توجهی بر جنبه های مختلف جامعه داشته باشند.

(۲) خروجی چنین پروژه هایی برای چه مواردی ممکن است کاربرد داشته باشد؟

صنعت املاک و مستغلات: شرکت ها و مشاوران املاک می توانند از این پیش بینی ها برای تعیین قیمت مناسب خانه ها برای فروش یا اجاره استفاده کنند. همچنین درک بهتر روندها و الگوهای قیمت در بازار مسکن یک منطقه و شناسایی عوامل موثر بر قیمت خانه.

تصمیم گیری خرید و فروش: به خریداران کمک می کند تا قیمت مناسب برای خرید خانه را تخمین بزنند. یا به فروشندگان کمک می کند تا قیمت مناسب برای فروش خانه را تعیین کنند.

سرمایه گذاری: سرمایه گذاران می توانند از این پیش بینی ها برای تصمیم گیری در مورد سرمایه گذاری در بازار مسکن استفاده کنند.

تحلیل بازار: دولت و سازمان های مرتبط می توانند از این پیش بینی ها برای تحلیل بازار مسکن و برنامه ریزی شهری جهت کمک به مسئولان در تصمیم گیری های مربوط به توسعه مسکن و زیرساخت ها استفاده کنند.

تصمیم گیری های مالی: افراد می توانند از این پیش بینی ها برای تصمیم گیری های مالی مانند خرید یا فروش خانه استفاده کنند.

پژوهش های علمی: پژوهشگران و دانشجویان می توانند از این پروژه برای انجام تحقیقات و پروژه های علمی در زمینه های تحلیل داده های مسکن استفاده کنند.

با توجه به این کاربردها، خروجی چنین پروژه هایی می تواند به بهبود تصمیم گیری ها، افزایش دقت پیش بینی ها و بهبود استفاده از داده ها کمک کند.

(۳) چه کسانی ممکن است به نتایج این پروژه علاقمند باشند؟ چرا؟

نتایج یک پروژه پیش بینی قیمت خانه با استفاده از الگوریتم رگرسیون خطی می تواند به افراد و سازمان های مختلفی کمک کند و علاقه مندی آن ها را جلب کند. افراد و سازمان های زیر ممکن است به نتایج این پروژه علاقه مند باشند:

مشاوران املاک و مستغلات: مشاوران املاک و مستغلات ممکن است به نتایج این پروژه علاقه‌مند باشند تا بتوانند قیمت‌های مناسب برای خرید و فروش خانه‌ها را تعیین کنند.

سرمایه‌گذاران: سرمایه‌گذاران ممکن است به نتایج این پروژه علاقه‌مند باشند تا بتوانند تصمیمات بهتری در مورد سرمایه‌گذاری در بازار مسکن اتخاذ کنند.

دولت و سازمان‌های شهرداری: دولت و سازمان‌های شهرداری ممکن است به نتایج این پروژه علاقه‌مند باشند تا بتوانند تحلیل‌های دقیق‌تری از بازار مسکن داشته باشند و برنامه‌ریزی‌های مناسبی را اجرا کنند.

تحقیق‌گران و دانشجویان: تحقیق‌گران و دانشجویان ممکن است به نتایج این پروژه علاقه‌مند باشند تا بتوانند از آن برای انجام تحقیقات و پروژه‌های علمی در زمینه‌ی تحلیل داده‌های مسکن استفاده کنند.

با توجه به اینکه نتایج این پروژه می‌تواند به بهبود تصمیم‌گیری‌ها، افزایش دقت پیش‌بینی‌ها و بهبود استفاده از داده‌ها کمک کند، افراد و سازمان‌های مختلفی ممکن است به نتایج این پروژه علاقه‌مند باشند.

درک داده :

(۱) داده‌ها از کجا بدست آمده‌اند و چگونه جمع‌آوری شده‌اند؟

داده‌های مورد استفاده در یک پروژه پیش‌بینی قیمت خانه ممکن است از منابع مختلفی بدست آمده باشند. برخی از روش‌های جمع‌آوری داده‌ها شامل موارد زیر می‌شود:

منابع آنلاین: داده‌های مربوط به قیمت خانه، متوسط درآمد ساکنان، متوسط عمر خانه‌ها و سایر ویژگی‌ها ممکن است از وبسایت‌های معتبر ملکی یا سایت‌های دولتی مربوط به مسکن جمع‌آوری شود.

منابع دولتی: دولت و سازمان‌های مرتبط معمولاً داده‌های مربوط به مسکن را جمع‌آوری و منتشر می‌کنند که می‌توان از آن‌ها برای تحلیل‌های مختلف استفاده کرد.

منابع خصوصی: شرکت‌ها و سازمان‌های خصوصی ممکن است داده‌های مربوط به خرید و فروش خانه‌ها را جمع‌آوری کرده و برای استفاده‌ی عمومی یا تحقیقاتی منتشر کنند.

نظرسنجی‌ها و مطالعات: برخی از داده‌ها از طریق نظرسنجی‌ها و مطالعات میدانی جمع‌آوری می‌شوند که اطلاعات دقیق‌تری از ویژگی‌های مسکن و قیمت‌ها ارائه می‌دهند.

بعد از جمع‌آوری داده‌ها، مراحل پیش‌پردازش اطلاعات مانند پاکسازی داده‌ها، تبدیل فرمت‌ها، جمع‌آوری داده‌های ناقص و غیرقابل اعتماد و حذف داده‌های تکراری انجام می‌شود تا داده‌های آماده برای تحلیل و استفاده در الگوریتم رگرسیون خطی باشند.

(۲) هر یک از متغیرها چه چیزی را اندازه‌گیری میکنند؟

پنج متغیر اول، متغیرهای مستقل هستند (X) و متغیر شماره شش یعنی "قیمت فروخته شده هر خانه" متغیر وابسته است (Y). براساس مقادیر X مقادیر Y مشخص میشوند یعنی X متغیر مستقل است و متغیر Y وابسته است. و متغیر address چون از نوع رشته‌ای بود و تأثیری در قیمت خانه نداشت، از محاسبات حذف گردید.

- **'Avg. Area Income':** Average. Income of residents of the city house is located in درآمد متوسط ساکنان منطقه.
- **'Avg. Area House Age':** Average Age of Houses in same city عمر متوسط خانه‌ها
- **'Avg. Area Number of Rooms':** Average Number of Rooms for Houses in same city متوسط تعداد اتاقهای هر خانه
- **'Avg. Area Number of Bedrooms':** Average Number of Bedrooms for Houses in same city متوسط تعداد اتاق خوابهای هر خانه
- **'Area Population':** Population of city house is located in جمعیت شهر
- **'Price':** Price that the house sold at قیمت فروخته شده هر خانه
- **'Address':** Address for the house آدرس

متوسط درآمد ساکنان منطقه: این متغیر میانگین درآمد یا سطح درآمد افراد ساکن در منطقه مورد مطالعه را اندازه گیری می کند که می تواند نشان دهنده قدرت مالی و توان خرید ساکنان باشد.

متوسط عمر خانه ها: این متغیر میانگین سن یا قدمت خانه های موجود در منطقه را اندازه گیری می کند. این متغیر می تواند نشان دهنده وضعیت کالبدی و فیزیکی خانه ها باشد.

متوسط تعداد اتاق های هر خانه: این متغیر میانگین تعداد اتاق های موجود در هر خانه را اندازه گیری می کند که می تواند نشان دهنده اندازه و متراژ خانه ها باشد.

متوسط تعداد اتاق خواب های هر خانه: این متغیر میانگین تعداد اتاق خواب موجود در هر خانه را اندازه گیری می کند و می تواند نشان دهنده اندازه و نوع خانه ها (مثلاً آپارتمان یا خانه مجزا) باشد.

جمعیت منطقه: این متغیر تعداد ساکنان یا جمعیت کل منطقه مورد مطالعه را اندازه گیری می کند و نشان دهنده تراکم جمعیتی در منطقه است. قیمت فروخته شده خانه: این متغیر قیمت نهایی خانه های فروخته شده در منطقه را اندازه گیری می کند و این متغیر هدف (target) یا خروجی مورد پیش بینی در پروژه است.

آدرس: این متغیر آدرس یا موقعیت جغرافیایی خانه های مورد مطالعه را اندازه گیری می کند. این متغیر ممکن است به عنوان یک ویژگی (feature) در پروژه استفاده شود.

۳) آیا ابهامی در تعاریف داده ها وجود دارد؟

پروژه خود، نیاز به توضیحات بیشتر یا تعیین دقیق تر متغیرها و ویژگی ها داشته باشید. برای مثال، ممکن است نیاز به تعریف دقیق تر متغیرهای وابسته و مستقل، روش های پیش پردازش داده، یا انتخاب ویژگی های مناسب برای مدل رگرسیون داشته باشید.

در فایل داده شده، داده خالی یا گمشده (missing value) و مقدار تکراری (duplicated) نداشتیم و همه ی متغیرها از نوع داده ای بودند و نیازی به تغییر نوع داده نداشتیم

با این حال با توجه به توضیحات ارائه شده، به نظر می رسد که تعاریف متغیرهای موجود در بانک داده فرضی به طور نسبتاً واضحی ارائه شده است. با این حال، ممکن است یک یا چند مورد از این متغیرها نیاز به توضیح بیشتر داشته باشند:

متوسط درآمد ساکنان منطقه: ممکن است نیاز به توضیح بیشتر در مورد چگونگی محاسبه این متوسط (مثلاً میانگین، میانه یا مد درآمد) باشد. همچنین واحد اندازه گیری درآمد (مثلاً تومان، دلار، یورو) باید مشخص شود.

متوسط عمر خانه ها: نیاز به توضیح در مورد اینکه آیا عمر خانه ها از زمان ساخت محاسبه شده است یا از زمان آخرین بازسازی.

آدرس: ممکن است نیاز به توضیح بیشتر در مورد چگونگی ثبت و فرمت آدرس خانه ها باشد (مثلاً استفاده از طول و عرض جغرافیایی، کد پستی یا ...).

در مجموع، به نظر می رسد که تعاریف ارائه شده برای اکثر متغیرها نسبتاً واضح هستند. اما برای رفع هرگونه ابهام احتمالی، ارائه توضیحات بیشتر در مورد نحوه محاسبه و ثبت داده ها می تواند مفید باشد.

۴) آیا ممکن است در اندازه گیری متغیرها و یا ثبت داده ها خطایی وجود داشته باشد؟

بله، احتمال وجود خطا در اندازه گیری متغیرها و ثبت داده ها در این نوع پروژه ها وجود دارد. برخی از مهم ترین منابع احتمالی خطا عبارتند از:

خطای اندازه گیری: برخی متغیرها مانند متوسط درآمد، عمر خانه ها و متراژ ممکن است به صورت نمونه گیری یا پیمایش محاسبه شوند. در این موارد خطای اندازه گیری وجود دارد. روش اندازه گیری و ابزارهای مورد استفاده نیز می تواند بر دقت داده ها تأثیر بگذارد.

خطای ثبت داده: ممکن است در ثبت و ورود داده ها به سیستم، اشتباهات تایپی یا خطاهای انسانی رخ دهد. روش های جمع آوری داده (مثل استفاده از پرسشنامه و مصاحبه) نیز ممکن است بر دقت داده ها تأثیر بگذارد.

خطای نمونه گیری: چنانچه داده ها تنها از نمونه ای از خانه ها جمع آوری شده باشند، ممکن است نماینده جامعه نباشند. انتخاب نمونه به شکل نادرست می تواند منجر به خطا در داده ها شود.

خطاهای سیستماتیک: وجود خطاهای سیستماتیک در داده ها (مثلاً به دلیل استفاده از منابع ناقص) می تواند به نتایج مدل پیش بینی آسیب برساند. بنابراین، شناسایی و کاهش منابع احتمالی خطا در داده ها از اهمیت بالایی برخوردار است و باید در طراحی و اجرای پروژه مورد توجه قرار گیرد.

۵) چه متغیرهای دیگری اگر وجود داشتند، میتوانست به حل مسئله کمک کند؟

در صورت در دسترس بودن برخی متغیرهای دیگر، می توان به بهبود دقت و کارایی مدل پیش بینی کمک کرد. برخی از این متغیرهای مفید تکمیلی عبارتند از:

ویژگی های فیزیکی خانه: متراژ زیربنا یا مساحت کل خانه - تعداد طبقات - وضعیت ساخت (نوسازی، بازسازی و...) - نوع ساختمان (آپارتمان، ویلایی و...) - امکانات و تجهیزات موجود در خانه

موقعیت و محیط خانه: فاصله از مراکز شهری و خدماتی - قرار گرفتن در محله خوب یا بد - دسترسی به حمل و نقل عمومی - میزان آلودگی محیطی

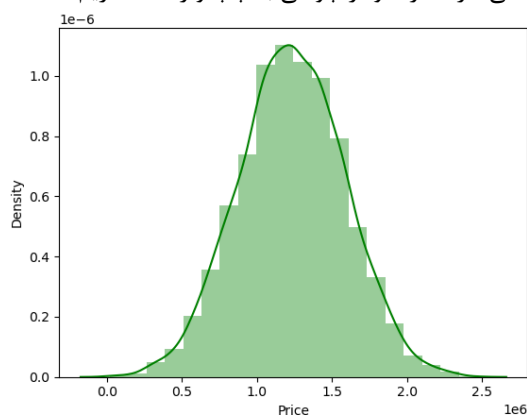
مشخصات منطقه: سطح درآمد متوسط ساکنان - امنیت و میزان جرایم - کیفیت مدارس و خدمات عمومی - میزان توسعه تجاری و خدماتی

عوامل اقتصادی: نرخ تورم و تغییرات قیمت - میزان دسترسی به تسهیلات مالی - وضعیت اقتصادی و رکود/رونق بازار

همچنین می توان به عواملی مانند: داشتن پارکینگ یا امکان پارک خودرو در معبر، اسکلت و سازه ضد زلزله، وجود آسانسور در ساختمان های چند طبقه و ... نیز توجه کرد زیرا در قیمت خانه تاثیر دارند.

۶) متغیرهای موجود از کدام نوعند (رشته ای - عددی) ؟ متغیرها از نوع عددی بودند و فقط سون آدرس از نوع رشته ای بود که در محاسبات لحاظ نگردید.

۷) خلاصه آماری متغیرهای موجود چیست؟ در تحلیل آماری از توابع pairplot و displot استفاده کردیم. برای تعیین نرمال بودن قیمت خانه از تابع distplot بهره بردیم. قیمت خانه ها از یک میلیون تا دو میلیون هستند و بیشترین قیمت خانه روی ۱.۵ میلیارد است و این مقدار نزدیک به میانه داده است بنابراین توزیع نرمالی دارد) در نمودار چولگی به چپ و راست نداریم



آماده سازی داده:

۱) آیا نیاز به در آمیختن داده ها است؟ اقدامات و نتایج گزارش شود.

باید داده های مرتبط شناسایی و جمع آوری شوند که در اینجا استفاده از چندین ستون ابتدایی به عنوان متغیر مستقل بود که آنها را با x مشخص نمودیم و همین طور قیمت را به عنوان متغیر وابسته (Y) معرفی کردیم. (بخش `# separating independence(x) and dependence(y) variables` در فایل HTML)

```
x=usahouse.iloc[:,5]
y=usahouse.iloc[:,2]
```

۲) آیا نیاز به پاکسازی داده است؟ اقدامات و نتایج گزارش شود.

از آنجایی که داده های گمشده نداشتیم، باید داده های پرت شناسایی و حذف شوند.

۳) آیا نیاز به تبدیل داده است؟ اقدامات و نتایج گزارش شود.

در اطلاعات داده شده مثلا تعداد اتاق خوابها ۳۰۹ آمده و جمعیت ۴۴ هزار و در نتیجه داده ها نرمال نیستند و خیلی پایین و بالا هستند، با هم همخوانی ندارند و اگر آن ها را به همین شکل باقی بگذاریم باعث ایجاد سوگیری می شوند بنابراین باید داده ها استاندارد شوند. که با استفاده از کتابخانه `sklearn.preprocessing` و تابع `standardscaler` داده ها را استاندارد و یکسان می کنیم. (بخش `#Standardize` در فایل HTML)

۴) آیا نیاز به کاهشی داده است؟ اقدامات و نتایج گزارش شود.

فقط ستون `Adress` از نوع `object` است که آنرا حذف میکنیم بقیه ستونها نوعشان `float 64` است که نیازی به تبدیل ندارند. با دستور `info()` اطلاعات کاملی از نوع و خالی بودن سطرها نشان داده میشود. که هیچ مقدار خالی یا گمشده ای در این دیتاها نداشتیم.

مدلسازی:

۱) روی داده های آموزش، الگوریتم رگرسیون ساخته شود. اقدامات و نتایج گزارش شود.

در بخش `Training a linear regression model` در فایل HTML همراه، یک مدل رگرسیون خطی روی داده های خود ایجاد کردیم. ابتدا متغیرهای مستقل (X) و متغیر وابسته یعنی (Y) را مشخص نموده تا بر اساس متغیرهای وابسته مقدار Y را پیش بینی کنیم. از آنجایی که الگوریتم ما از نوع با نظارت است باید X و Y را به مدل بدهیم تا یاد بگیرد و سپس از روی آن برای ما پیش بینی انجام دهد تا مقادیر B_0 و B_1 را برای داشتن بهترین مدل برای ما به دست بیاورد (B_0 ضریب و B_1 شیب خط است)

$$Y_i = B_0 + B_1 x_i$$

و بین مقادیر مشاهده شده با مقادیر پیش بینی شده تفاوتی وجود دارد که خطا نامیده می شود و هر چه این خطا کمتر باشد، مدل ما بهتر است.

برای اینکه مطمئن شویم مدل خوبی انتخاب کرده ایم، داده ها را به دو بخش آموزش و آزمایش تقسیم می کنیم که در اینجا ۳۰۰۰ رکورد برای آموزش و ۲۰۰۰ رکورد برای آزمایش تعریف شدند.

بعد از استاندارد سازی داده ها، در بخش `Creating and training the model` از مدل رگرسیون خطی استفاده کرده و با تابع `Fit` به آن آموزش دادیم.

ارزیابی:

۱) مدل های ارائه شده، روی داده های آزمایش با استفاده از شاخص های ارزیابی رگرسیون خطی در یادگیری ماشین ارزیابی

شوند. اقدامات و نتایج گزارش شود.

در بخش `find model coefficients` مقدار B_0 و مقادیر B_i را برای ما پیدا کرده و بعد در بخش `prediction` می خواهیم که برای ما مقادیر Y را پیش بینی کند و سپس با رسم نمودار متوجه می شویم که پیش بینی خیلی خوب بوده چون مبتنی بر زاویه خطی ۴۵ درجه است و مقادیر پیش بینی شده با مقادیر واقعی به هم بسیار نزدیکی هستند در نتیجه این مدل، مدل خوبی است.

۲) چه پیشنهاداتی دارید تا نتایج در محیط واقعی، آزمایش گردد؟

برای آزمایش نتایج پروژه خود در محیط واقعی و اعتبارسنجی آن‌ها، می‌توانید از روش‌ها و راهکارهای زیر استفاده کنید:

آزمون در محیط واقعی: اجرای آزمایش‌های واقعی با استفاده از مدل پیش‌بینی شده بر روی داده‌های جدید و واقعی. این کار به شما کمک می‌کند تا نتایج مدل خود را در شرایط واقعی ارزیابی کنید.

مقایسه نتایج: مقایسه نتایج پیش‌بینی مدل با وضعیت واقعی و مشاهده‌ی تطابق یا عدم تطابق آن‌ها. این مقایسه می‌تواند به شما کمک کند تا دقت و کارایی مدل خود را ارزیابی کنید.

استفاده از معیارهای ارزیابی: استفاده از معیارهای ارزیابی مانند میانگین مربعات خطا (Mean Squared Error) یا ضریب تعیین (Coefficient of Determination) برای اندازه‌گیری دقت مدل در محیط واقعی.

تحلیل حساسیت: بررسی تأثیر تغییرات در ورودی‌ها بر خروجی مدل و ارزیابی عملکرد آن در شرایط مختلف. این تحلیل می‌تواند به شما کمک کند تا درک بهتری از رفتار مدل در مقابل تغییرات داشته باشید.

استفاده از مدل‌های مقایسه‌ای: مقایسه نتایج مدل خود با مدل‌های مشابه دیگری که برای پیش‌بینی قیمت خانه استفاده می‌شوند. این مقایسه می‌تواند به شما کمک کند تا عملکرد مدل خود را در مقایسه با رقبا ارزیابی کنید.

با انجام این مراحل و اعتبارسنجی دقیق مدل خود در محیط واقعی، می‌توانید اطمینان حاصل کنید که نتایج پروژه شما به خوبی عملکرد می‌کنند و قابل اعتماد هستند.

استقرار:

حال اگر بخواهید چنین الگوریتمی را در مقیاس صنعتی توسعه دهید، به این فکر کنید با چه چالش‌هایی مواجه خواهید شد و برای آن چه راهکارهایی دارید. موارد زیر را گزارش کنید:

۱) چالش‌های توسعه الگوریتم را بررسی کنید.

توسعه یک الگوریتم رگرسیون خطی برای پیش‌بینی قیمت خانه در مقیاس صنعتی با چالش‌های زیر مواجه خواهد شد:

مقیاس پذیری: تطبیق الگوریتم برای پردازش حجم بزرگی از داده‌ها و اجرای آن در مقیاس صنعتی می‌تواند چالش‌هایی ایجاد کند. بهینه‌سازی الگوریتم برای اجرا در محیط‌های پردازش موازی و توزیع شده می‌تواند لازم باشد.

مدیریت داده: مدیریت و ذخیره‌سازی داده‌های حجیم می‌تواند یک چالش مهم باشد. انتقال، ذخیره و بازیابی داده‌ها به صورت کارآمد و امن برای عملکرد بهینه الگوریتم ضروری است.

پیش‌پردازش داده: پیش‌پردازش داده‌ها به صورت موثر و دقیق می‌تواند یک چالش باشد. تمیز کردن داده‌ها، تبدیل فرمت‌ها، حذف داده‌های ناقص و تبدیل داده‌های اولیه به ورودی‌های قابل استفاده برای الگوریتم می‌تواند زمان‌بر و پیچیده باشد.

تطبیق با محیط‌های مختلف: الگوریتم باید قابلیت تطبیق با محیط‌های مختلف و شرایط مختلف را داشته باشد. این شامل توانایی عملکرد در مقیاس‌های مختلف، محیط‌های متفاوت و داده‌های متغیر است.

بهینه‌سازی و بهره‌وری: بهینه‌سازی الگوریتم برای اجرا به صورت سریع و با بهره‌وری بالا می‌تواند یک چالش مهم باشد. استفاده از الگوریتم‌های بهینه و بهره‌ور می‌تواند در افزایش کارایی و کاهش زمان اجرا کمک کند.

جمع‌آوری داده‌ها و پاکسازی: به دست آوردن و نگهداری اطلاعات دقیق، مرتبط و تمیز می‌تواند بسیار دشوار باشد.

حفظ عملکرد در طول زمان: همانطور که تغییرات بازار و داده‌های بیشتر جمع‌آوری می‌شود، مدل باید به روز شود و برای حفظ عملکرد، آموزش داده شود.

برای مواجهه با این چالش‌ها، می‌توان از روش‌ها و راهکارهایی مانند بهینه‌سازی الگوریتم، استفاده از فناوری‌های پیشرفته مانند پردازش موازی و توزیع شده، استفاده از سیستم‌های مدیریت داده قدرتمند، انجام پیش‌پردازش داده‌ها با دقت و دقیق، و توسعه راه‌حل‌های انعطاف‌پذیر برای تطبیق با محیط‌های مختلف استفاده کرد.

(۲) چه راهکارهایی برای حل آن ها دارید؟

راهکارهای ممکن برای حل چالش‌های توسعه الگوریتم در مقیاس صنعتی:

مقیاس پذیری:

- استفاده از فناوری‌های پردازش موازی و توزیع شده برای افزایش سرعت و کارایی الگوریتم.
- بهینه‌سازی الگوریتم برای محیط‌های مختلف و اجرا در مقیاس‌های مختلف.

مدیریت داده:

- استفاده از سیستم‌های مدیریت داده قدرتمند مانند پایگاه داده‌های توزیع شده برای ذخیره و مدیریت داده‌ها.
- استفاده از فناوری‌های ابری برای ذخیره و دسترسی به داده‌ها به صورت ایمن و مؤثر.

پیش‌پردازش داده:

- اتوماسیون فرآیند پیش‌پردازش داده‌ها با استفاده از ابزارها و روش‌های خودکار.
- استفاده از الگوریتم‌های پیش‌پردازش داده‌ها برای تمیز کردن و تبدیل داده‌ها به ورودی‌های قابل استفاده برای الگوریتم.

تطبیق با محیط‌های مختلف:

- طراحی الگوریتم به صورت انعطاف‌پذیر و قابل تنظیم برای تطبیق با شرایط و محیط‌های مختلف.
- استفاده از معماری میکروسرویس‌ها برای اجرای الگوریتم در محیط‌های مختلف و به صورت مستقل از یکدیگر.

بهینه‌سازی و بهره‌وری:

- استفاده از الگوریتم‌های بهینه و بهره‌ور برای کاهش زمان اجرا و افزایش کارایی.
 - استفاده از تکنیک‌های بهینه‌سازی مانند تکنیک‌های تکراری، کاهش ابعاد داده و بهینه‌سازی مسائل بهینه‌سازی.
- با اجرای این راهکارها و استفاده از فناوری‌های پیشرفته، می‌توانید بهبود عملکرد الگوریتم خود در محیط صنعتی را تضمین کنید و با چالش‌های توسعه مواجه شده روبرو شوید.

(۳) چه ملزوماتی برای ارائه آن راهکارها نیاز دارید؟

داده‌های بزرگ، مرتبط و با کیفیت بالا: این مدل باید به یک مجموعه داده بزرگ و متنوع دسترسی داشته باشد که به طور دقیق نشان دهنده بازار املاک و مستغلات و عواملی است که بر قیمت خانه‌ها تاثیر می‌گذارد. پردازش مقادیر زیادی از داده‌ها نیازمند قدرت محاسباتی قابل توجه است. راه حل‌های مبتنی بر ابر یا سیستم‌های محاسباتی توزیع شده ممکن است برای مقیاس عملکرد الگوریتم ضروری باشد.

تخصص در علوم داده و یادگیری ماشین: یک تیم ماهر از دانشمندان داده و مهندسان یادگیری ماشین باید برای طراحی، پیاده‌سازی و حفظ الگوریتم در دسترس باشند

ادغام و استقرار مستمر: یک فرآیند خودکار باید به صورت دوره‌ای به روز شود و مدل را با داده‌های جدید بازگزداند و اطمینان حاصل شود که الگوریتم دقیق و به روز مانده است

مدیریت داده‌های کارآمد: جمع‌آوری داده‌های قوی، ذخیره‌سازی و مکانیسم‌های بازیابی برای اطمینان از دسترسی به موقع به داده‌های مربوطه و دقیق، ضروری است

نتیجه گیری:

(۱) انجام این پروژه چه یادگیری برای شما داشت؟

- اهمیت دسترسی به داده‌های با کیفیت بالا: به منظور توسعه مدل‌های پیش‌بینی دقیق، نیاز به دسترسی به مجموعه داده‌های گسترده و متنوع است که به طور دقیق نشان دهنده اطلاعات مورد نیاز باشد
- نیاز به یادگیری و بهینه‌سازی مداوم

(۲) با چه چالشهایی مواجه شدید؟ چگونه آنها را حل کردید؟

عدم آشنایی صحیح با رگرسیون خطی در تحلیل و پیش بینی (راه حل : مطالعه درباره موضوع)
عدم شناخت بازار مسکن، عوامل تاثیر گذار بر قیمت خانه ها و چالش های جمع آوری اطلاعات (راه حل : تحقیق و جستجو در اینترنت و پرسش از افراد مختلف)