تحلیل جامع بازار املاک در پلتفرم دیوار

نام موسسه: مدرسه پردازش و تحلیل داده دقیقه

تهیه‌کنندگان: کورش خدائی نمین/ زینب السادات حسینی / ایلیا طاهری‌نیا

معرفی مجموعه داده

+۱M

آگهی املاک

بیش از یک میلیون آگهی معتبر از پلتفرم دیوار

۶۰

ویژگی

شصت ستون ویژگی شامل اطلاعات مالی، فیزیکی و مکانی

۱۶

دسته‌بندی

شانزده نوع ملک از آپارتمان تا ویلا و املاک تجاری

این مجموعه داده بازتابی زنده از رفتار اقتصادی، قیمت‌گذاری، ترجیحات کاربران و پویایی‌های بازار مسکن در ایران است.

ساختار داده‌ها

مشخصات فیزیکی و مالی

- ابعاد ملک (متراژ زمین و بنا)
- قیمت، اجاره، رهن و ودیعه
- جزئیات ساختمان (طبقه، اتاق، سال ساخت)

دسته‌بندی و موقعیت

- نوع معامله و ملک (اجاره، فروش، آپارتمان، ویلا)
- موقعیت جغرافیایی (شهر، محله، مختصات)
- اطلاعات آگهی‌دهنده و زمان ثبت

امکانات و ویژگی‌های املاک

امکانات پایه

آسانسور، پارکینگ، انباری و بالکن

انشعابات اصلی

آب، برق، گاز و سیستم‌های گرمایشی و سرمایشی

امکانات لوکس

استخر، جکوزی، سونا و نگهبانی

تحلیل اکتشافی: شناخت اولیه

در نخستین گام‌های اکتشافی، ساختار داده‌ها، مقادیر گمشده و نوع داده‌ها بررسی شد. داده‌ها مخلوطی از رشته‌های فارسی و انگلیسی، اعداد با فرمت‌های مختلف و کاراکترهای خاص بودند.

مقادیر گمشده

درصد بالایی از ستون‌های مرتبط با مشخصات کمی ملک دارای مقادیر خالی بودند

نقاط پرت

وجود مقادیر غیرواقعی شدید در متغیرهای قیمت و ابعاد

استانداردسازی

نیاز به تبدیل فرمت نوشتاری و یکپارچه‌سازی داده‌ها

دسته‌بندی ویژگی‌ها

۰۱

متغیرهای عددی پیوسته

۱۶ ستون شامل قیمت، اجاره، رهن، متراژ و مختصات جغرافیایی

۰۲

متغیرهای عددی گسسته

۷ ستون شامل تعداد اتاق، طبقات، سال ساخت و ظرفیت

۰۳

متغیرهای بولین

۱۷ ستون نشان‌دهنده وجود یا عدم وجود امکانات

۰۴

متغیرهای دسته‌بندی

۱۷ ستون شامل نوع ملک، محله، جهت ساختمان و جنس کف

۰۵

متغیرهای متنی

۳ ستون شامل عنوان، توضیحات و تاریخ ثبت آگهی

کشف روابط: همبستگی با اجاره

همبستگی‌های ضعیف

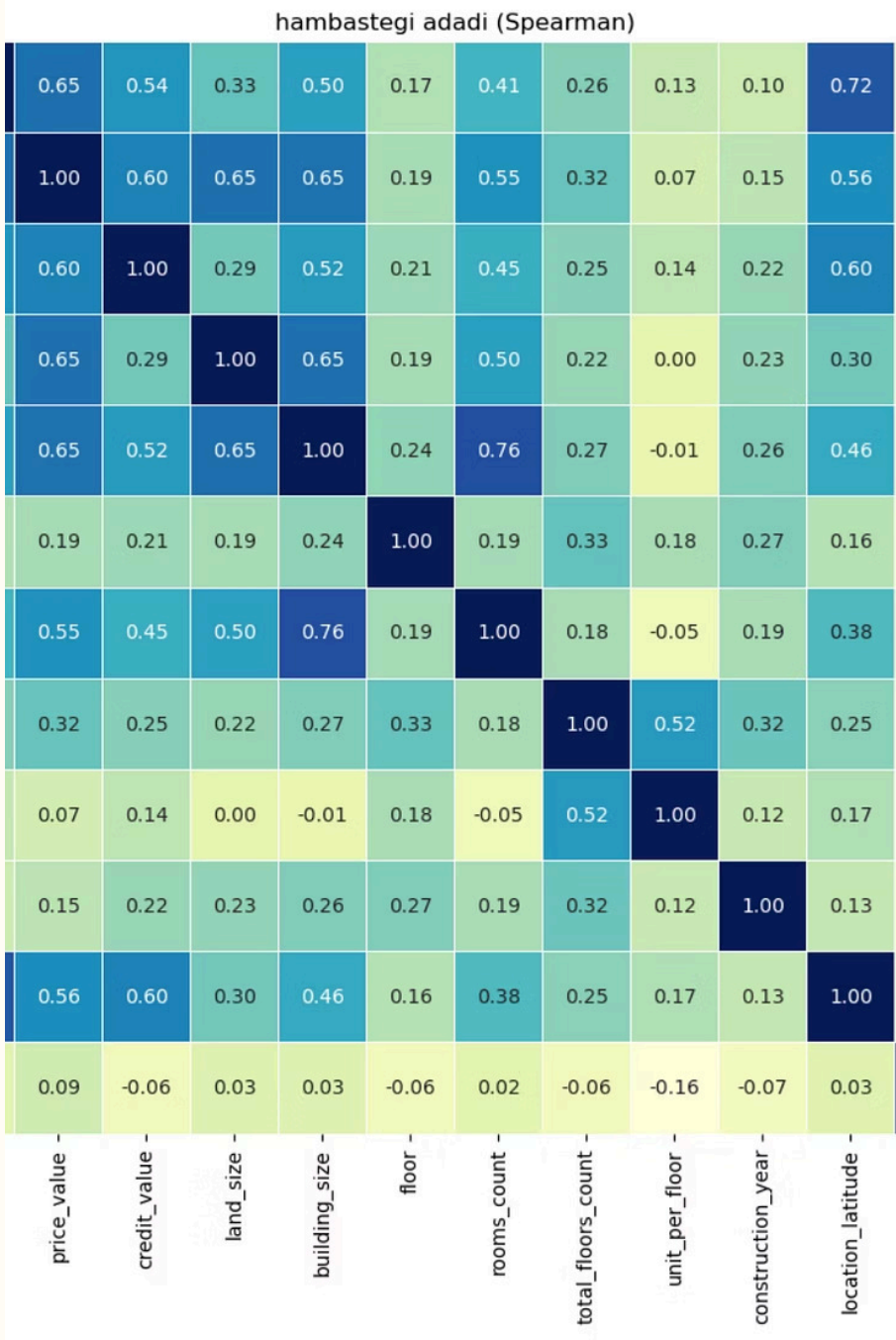
- floor: ۰.۱۷
- construction_year: ۰.۱۰
- location_longitude: ۰.۰۲

همبستگی‌های متوسط

- building_size: ۰.۵۰
- rooms_count: ۰.۴۱
- land_size: ۰.۳۳

قوی‌ترین همبستگی‌ها

- location_latitude: ۰.۷۲
- price_value: ۰.۶۵
- credit_value: ۰.۵۴



کشف روابط: همبستگی با قیمت

همبستگی خوب

- location_latitude: ۰.۵۶
- rooms_count: ۰.۵۵

قوی‌ترین همبستگی‌ها

- rent_value: ۰.۶۵
- land_size: ۰.۶۵
- building_size: ۰.۶۵
- credit_value: ۰.۶۰

پاکسازی داده‌ها: مدیریت تکرار

در بررسی داده‌های تکراری، چالش اصلی تشخیص املاک واقعاً یکسان از آگهی‌های مجدد بود. استراتژی تفکیک داده‌ها به دو دسته اتخاذ شد:

۱

داده‌های باکیفیت

رکوردهایی با مقادیر کامل در ستون‌های کلیدی مانند متراژ و قیمت

۲

شناسایی تکراری

اجرای الگوریتم فقط روی داده‌های باکیفیت

۳

حذف داده های تکراری

۸,۷۹۰ رکورد تکراری از میان ۹۸۱,۶۷۵ رکورد شناسایی و حذف شد

استانداردسازی و یکپارچه سازی

1

تبدیل اعداد

پاکسازی و تبدیل اعداد فارسی، حروف و علائم غیرعددی به فرمت استاندارد

دسته‌بندی

تبدیل ستون‌های دسته‌بندی به نوع categorical با مقدار missing برای NaNها

مقادیر بولین

یکپارچه‌سازی تمام خانه‌ها با مقادیر True یا False

خوشه بندی CLUSTERING: هدف و روش

هدف اصلی، گروه‌بندی املاک بر اساس ویژگی‌های عددی و نوعی به گونه‌ای که بیشترین شباهت ساختاری و رفتاری را داشته باشند.

۱

پیش‌پردازش

مدیریت مقادیر خالی، حذف پرت‌ها و استانداردسازی

۲

K–Means

اجرای الگوریتم و یافتن تعداد بهینه خوشه‌ها

۳

کاهش ابعاد

استفاده از PCA برای ارزیابی پایداری

۴

تحلیل نهایی

بررسی فضاهای ۱۱ و ۸ بعدی

مدیریت مقادیر گمشده

مقادیر گمشده براساس نوع معامله (cat۲_slug) بررسی شد. خالی بودن مقادیر فروش در معاملات اجاره بدیهی و اجتناب‌ناپذیر است.

حذف ستون‌های ضعیف

حذف ویژگی‌های کم‌ارزش

جایگزینی داده صحیح

جایگزینی مقادیر گم‌شده براساس تحلیل آماری

مدیریت مقادیر پرت

۱

شناسایی مقادیر غیرواقعی

قیمت‌های ۹۹۹ تریلیون تومان و متراژهای ۱۰ میلیون متر مربع

۲

حذف رکوردهای پرت

حذف آگهی‌های با داده‌های کاملاً غیرمنطقی

۳

تبدیل لگاریتمی

اعمال تبدیل log برای کاهش چولگی توزیع

۴

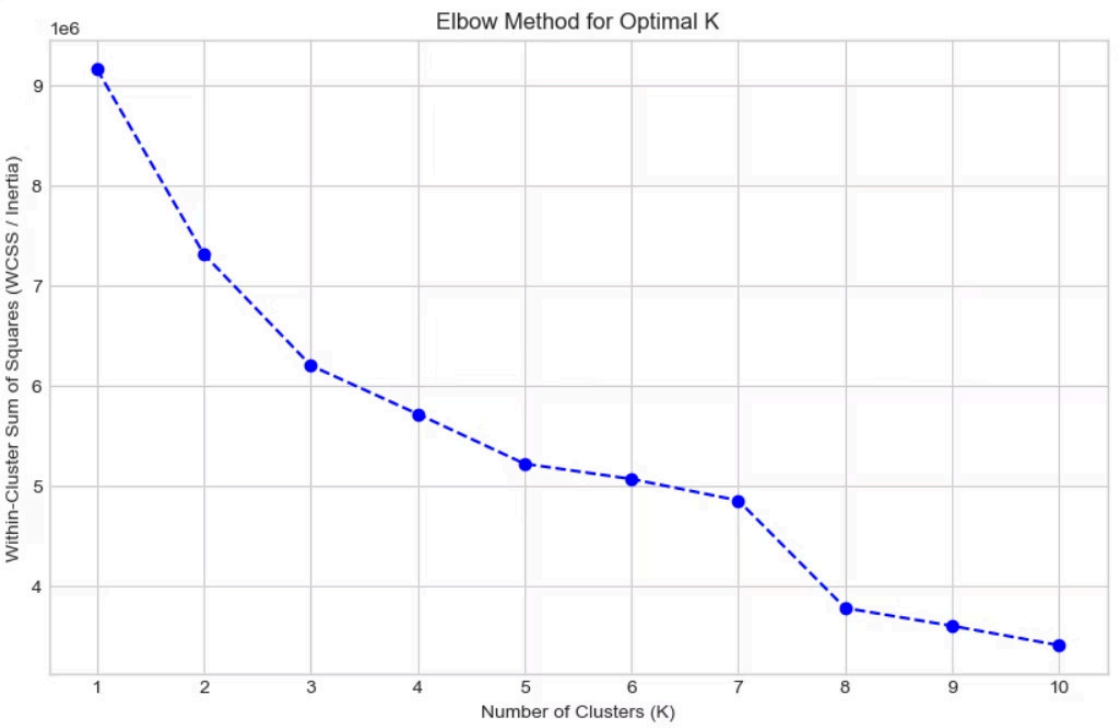
استفاده از چارک‌ها

محاسبه آستانه براساس IQR و توزیع واقعی داده‌ها

۵

تفکیک براساس نوع

اجرای مجدد براساس نوع ملک برای دقت بیشتر



پیدا کردن بهترین تعداد خوشه‌ها

۱ محاسبه WCSS

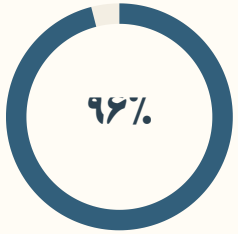
۲ رسم نمودار Elbow

۳ شناسایی نقطه بهینه K=۸

نتایج خوشه بندی

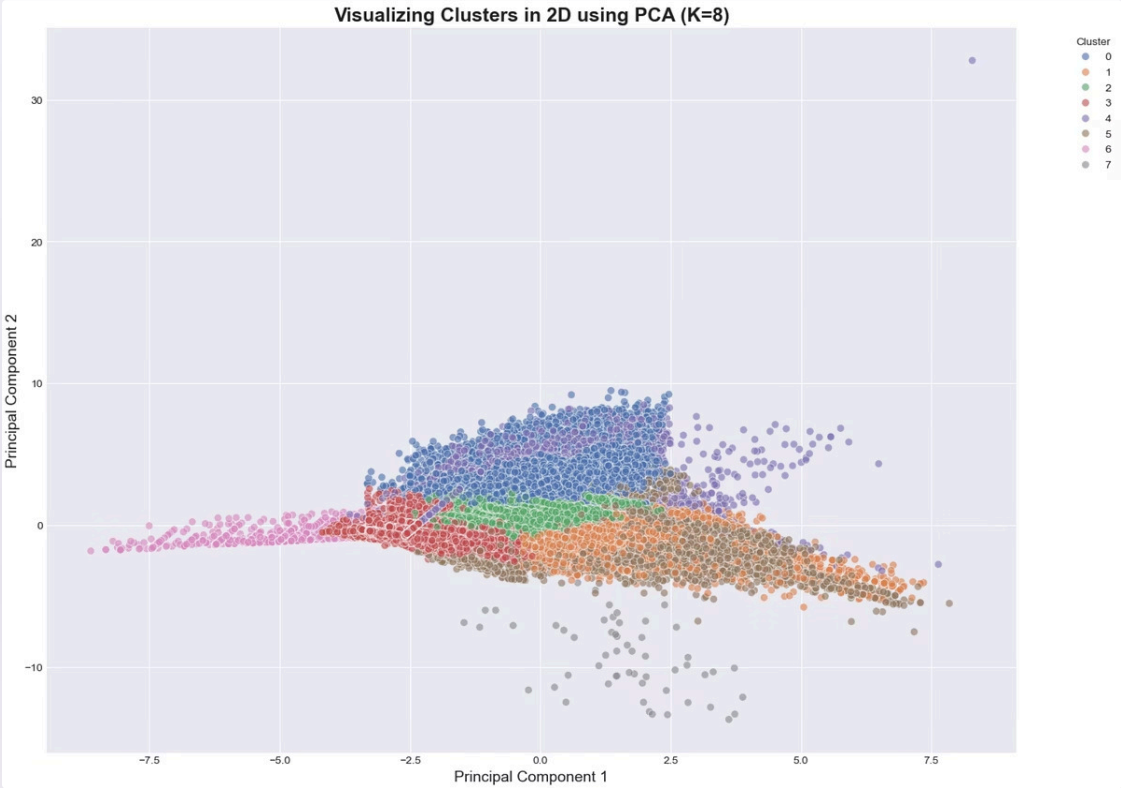
واریانس حفظ شده

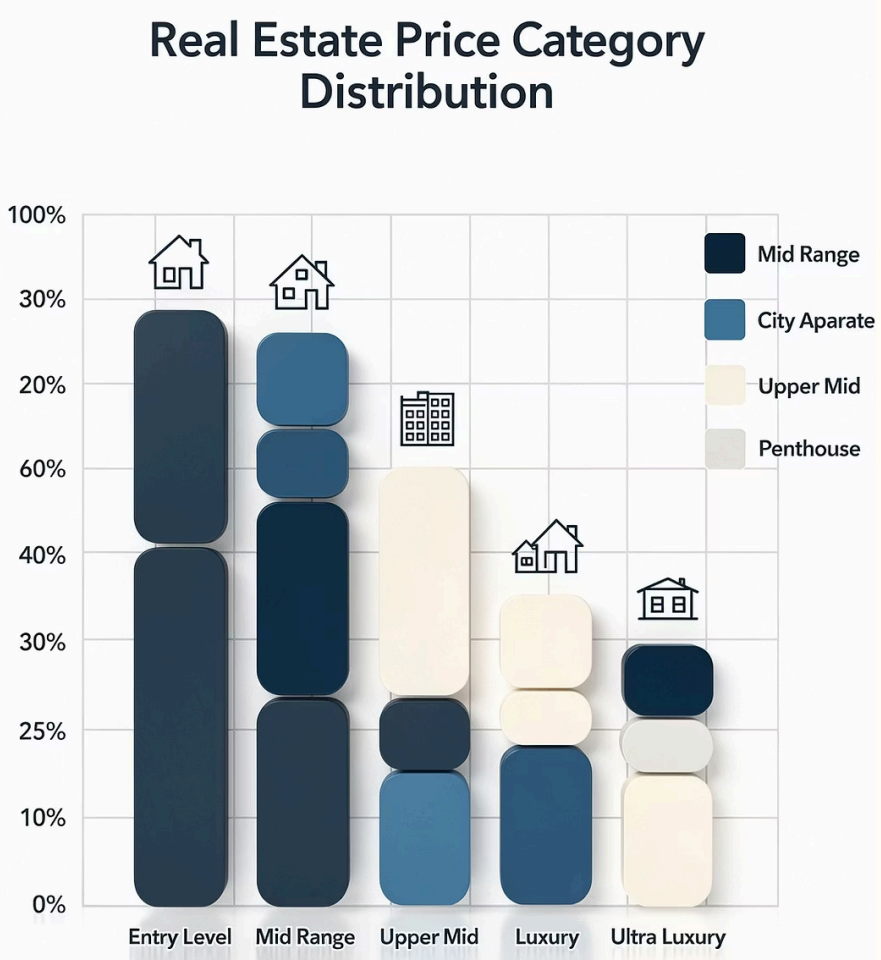
واریانس حفظ شده با ۸ مولفه اصلی در PCA



پایداری خوشه

حداقل پایداری خوشه در فضاهای مختلف





مدلسازی قیمت: آماده‌سازی

پیش از مدل‌سازی، یک مرحله پاکسازی جامع انجام شد تا داده‌ها به شکلی پایدار و قابل‌تحلیل آماده شوند.

پاکسازی قیمت‌ها

حذف داده‌های نادرست و پراکندگی

اصلاح ابعاد

همسان‌سازی و مقیاس‌بندی ویژگی‌ها

تکمیل ویژگی‌ها

پرکردن مقادیر گمشده و تزریق اطلاعات

مهندسی ویژگی

سن بنا

محاسبه سن ملک با کسر سال ساخت از ۱۴۰۳ برای مدل‌سازی فرسودگی

اثر طبقه و آسانسور

ویژگی ترکیبی با وزن ۱.۵ برای آسانسور و ۰.۷ بدون آسانسور

میانگین قیمت محله

محاسبه میانه قیمت در هر محله برای لحاظ کردن تأثیر منطقه

مقایسه مدل‌های رگرسیون: مراحل و نتایج

مرحله اول: مدل‌های پایه (بدون تنظیم)

KNN Regressor	۰.۸۳	بهترین عملکرد در تنظیم اولیه با وزن distance
Linear Regression	۰.۷۷	عملکرد قابل قبول با سادگی بالا
Decision Tree	۰.۷۶	نزدیک به رگرسیون خطی، نیاز به تنظیم
MLP (پایه)	منفی	عملکرد بسیار ضعیف، بدتر از میانگین

اثر تاثیر فراپارامترها در MLP

بهترین تنظیمات:

- معماری: یک لایه پنهان با ۶۴ نورون
- فعال‌سازی: ReLU
- حداکثر تکرار: حدود ۲۰۰

نتیجه: R^2 همچنان منفی (عملکرد ضعیف)

دلایل ضعف MLP

هم‌خطی و هم‌پوشانی زیاد ویژگی‌ها

معماری ساده نسبت به پیچیدگی الگوها

تعداد زیاد ویژگی‌های تنک (OneHot)

حتی با تغییرات متوالی در معماری، MLP نتوانست به دقت KNN یا Decision Tree نزدیک شود

دسته‌بندی قیمتی املاک

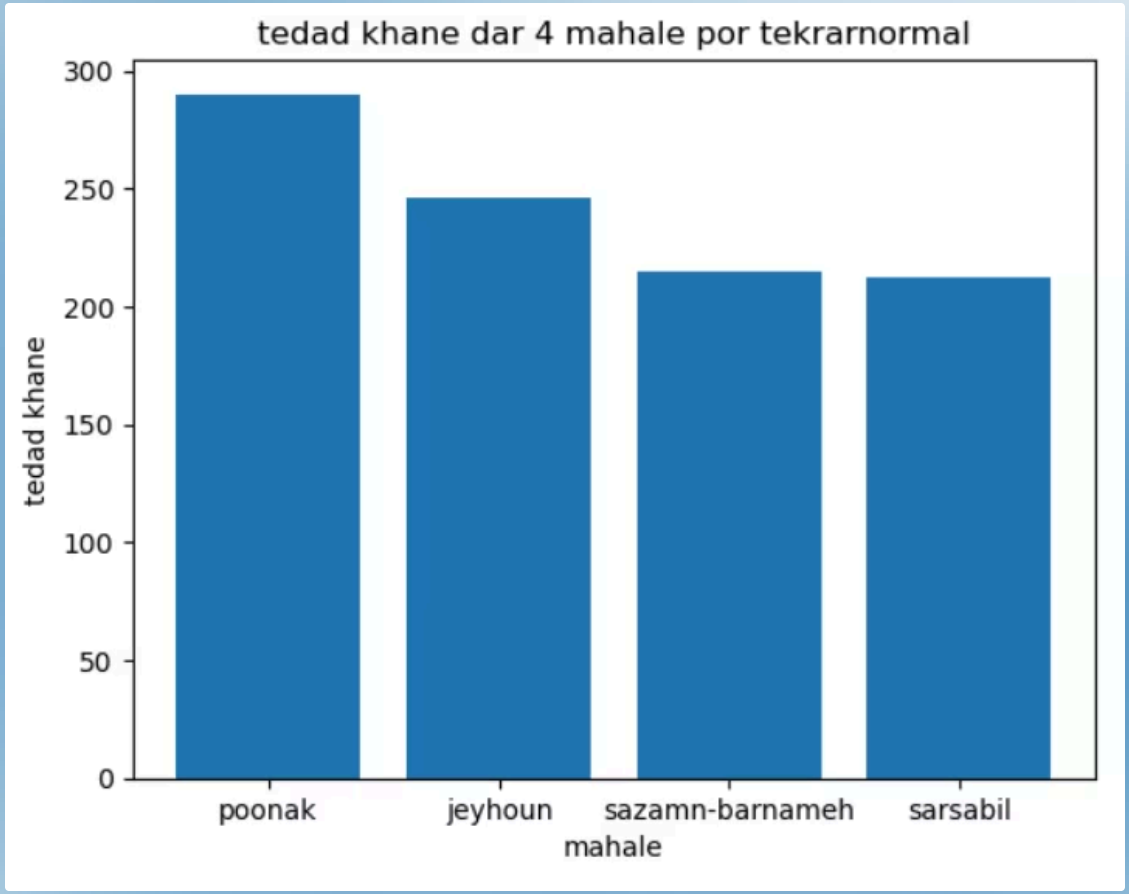
۹K
قیمت مناسب

۲K
گران

۱.۸K
ارزان

تحلیل مکانی دسته‌های قیمتی

بررسی توزیع جغرافیایی نشان داد که محله‌هایی مانند پونک، جیحون و سازمان برنامه بیشترین تعداد آگهی با قیمت مناسب را دارند.



قیمت مناسب

تعادل نسبی عرضه و تقاضا در مناطقی مانند پونک و جیحون

املاک گران

تمرکز در پونک، چیتگر-لیک و سرسبیل با قیمت ۲۰-۳۰٪ بالاتر

فرصت‌های خرید

املاک ارزان با قیمت نصف سطح مناسب، مناسب برای سرمایه‌گذاری

بارگذاری داده و بررسی اولیه برای دسته بندی متنی

عملیات بر روی خروجی مرحله قبل که حاوی متون پاکسازی، ریشه‌یابی و توکن‌سازی شده در ستون clean_text بود انجام شد.

رکورد کل

کلاس مختلف در cat۳_slug

رکورد غیرتهی در user_type

نمونه‌برداری تصادفی برای مدل‌سازی

استخراج ویژگی متنی و برداری‌سازی

مدل‌های مبتنی بر شمارش

- Bag of Words
- TF-IDF

مدل مبتنی بر تعبیه (Embedding)

- Word۲Vec

رویکرد یادگیری عمیق

- Tokenizer

مدل‌سازی دسته‌بندی cat۳_slug

برای مقایسه عملکرد رویکردهای مختلف، سه مدل دسته‌بندی بر روی داده‌های برداری شده آموزش داده شدند:

رگرسیون لجستیک

جنگل تصادفی

بیز ساده چندجمله‌ای

نتایج مدل‌های کلاسیک

مقایسه عملکرد مدل‌های مختلف نشان داد که Logistic Regression با TF-IDF بهترین نتایج را ارائه داد:

مدل (Model)	روش برداری‌سازی (Vectorization)	دقت (Accuracy)	F۱ وزن‌دار (Weighted F۱)
Logistic Regression	BoW	۰.۸۲۵	۰.۸۲۳
RandomForest	BoW	۰.۴۹۵	۰.۴۱۰
RandomForest+	BoW	۰.۶۹۸	۰.۷۱۶
NaiveBayes	BoW	۰.۷۴۹	۰.۷۵۵
Logistic Regression	TF-IDF	۰.۸۲۹	۰.۸۲۶
RandomForest	TF-IDF	۰.۴۹۴	۰.۴۱۳
RandomForest+	TF-IDF	۰.۶۹۸	۰.۷۱۶
NaiveBayes	TF-IDF	۰.۷۲۹	۰.۷۱۵

نتایج Word۲Vec: Accuracy ≈ ۰.۷۷۸, F۱ ≈ ۰.۷۷۴

بهترین مدل کلاسیک: Logistic Regression + TF-IDF با دقت ۸۲.۹٪ و F۱ وزن‌دار ۸۲.۶٪

معماری شبکه LSTM برای cat۳_slug

برای بهره‌گیری از اطلاعات ترتیبی و معنایی دقیق‌تر در متن، یک مدل یادگیری عمیق مبتنی بر LSTM طراحی شد:

لایه Embedding

تبدیل توکن‌ها به بردارهای ۱۰۰ بعدی

لایه LSTM

پردازش توالی و حفظ وابستگی‌های زمانی

لایه Dense + Dropout

استخراج ویژگی و جلوگیری از بیش‌برازش

لایه خروجی Softmax

دسته‌بندی به ۱۶ کلاس

جزئیات پیکربندی:

- **تابع زیان:** sparse_categorical_crossentropy
- **بهینه‌ساز:** Adam
- **اپوک ها:** ۱۰
- **اندازه بَچ:** ۳۲

نتایج و تحلیل آموزش LSTM

مدل LSTM توانست به دقت قابل توجه ۸۴٪ روی مجموعه آزمون دست یابد. تحلیل روند آموزش نشانه‌هایی از بیش‌برازش را نشان داد:

مشاهدات کلیدی

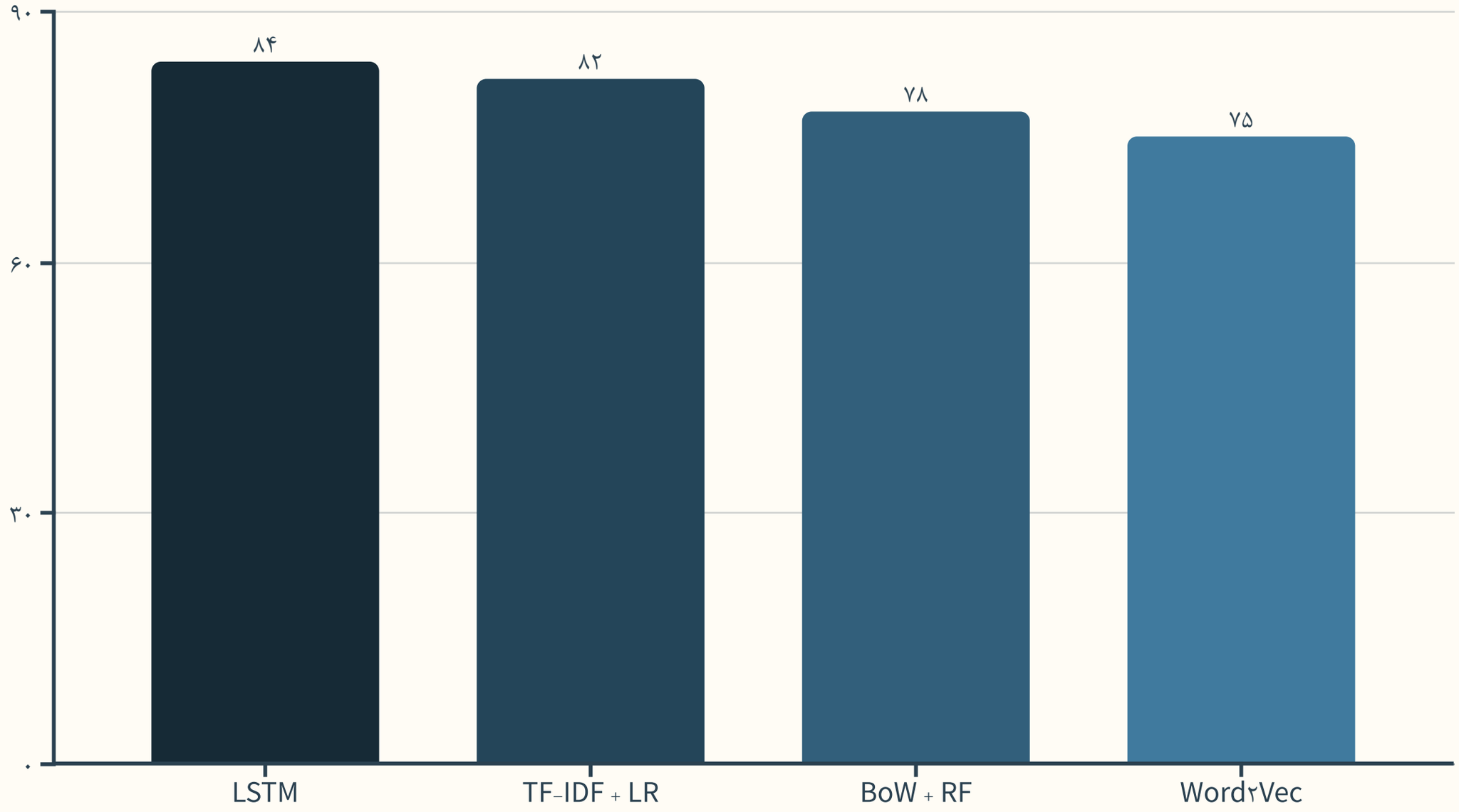
- بهترین عملکرد: epoch ۳-۵
- نشانه بیش‌برازش پس از epoch ۵
- تغیرات val_loss روند صعودی از ۰.۴۴ به ۰.۵۴
- توصیه: توقف زودهنگام در epoch ۳-۵

عملکرد مدل

- Accuracy ۰.۸۹ :آموزشی (epoch ۱۰)
- Validation Accuracy: ۰.۸۴
- Loss ۰.۲۸ :آموزشی
- Validation Loss: ۰.۵۴

دقت نهایی ۸۴٪ برای دسته‌بندی ۱۶ کلاسه بسیار رضایت‌بخش است (دقت تصادفی: ۶.۲۵٪)

مقایسه روش‌های برداری‌سازی



مقایسه روش‌های مختلف نشان داد که LSTM بهترین عملکرد را دارد، اما TF–IDF با Logistic Regression نیز با دقت ۸۲٪ گزینه مناسبی با هزینه محاسباتی کمتر است.

پیش‌بینی نوع کاربری (user_type)

با توجه به حجم بالای داده‌های گمشده در ستون user_type (تنها ۲۸۸,۸۷۷ از ۹۹۹,۹۴۵ رکورد)، یک مدل LSTM برای پیش‌بینی این مقادیر طراحی شد.

چالش: عدم توازن کلاس

- عدم توازن شدید کلاس‌ها
- مشاور املاک: **۱۰۲,۰۴۳** نمونه
- شخصی: **۱۳,۳۴۸** نمونه
- نسبت: **۷.۶ به ۱**

راه حل: وزن‌دهی کلاس

- وزن‌دهی کلاس (class weighting)
- وزن کلاس اقلیت (شخصی): **۴.۳۲**
- وزن کلاس اکثریت (مشاور): **۰.۵۷**
- معماری LSTM با **sigmoid**

نتایج: عملکرد مدل

- Accuracy: ۸۹٪
- Weighted F۱: ۰.۹۰
- Recall ۸۱ :کلاس شخصی٪
- Precision ۹۷ :مشاور املاک٪

بیش از ۷۰۰,۰۰۰ مقدار گمشده با موفقیت پیش‌بینی شد

دستاوردهای کلیدی پروژه

پروژه حاضر با هدف تحلیل جامع آگهی‌های ملکی تهران از پلتفرم دیوار، و با تمرکز بر استخراج الگوهای پنهان در بازار مسکن و ساخت مدل‌های قابل‌اعتماد برای پیش‌بینی قیمت و دسته‌بندی داده‌ها، طراحی و اجرا گردید.

پاکسازی و مهندسی داده

ایجاد مجموعه داده منسجم با مدیریت تکرارها، مقادیر پرت و گمشده

خوشه‌بندی املاک

شناسایی ۸ خوشه پایدار (۹۶-۱۰۰٪) با پوشش ۹۰.۲۲٪ واریانس

مدل‌سازی قیمت

پیش‌بینی دقیق قیمت و دسته‌بندی سه‌سطحی ارزش املاک

تحلیل متنی NLP

دقت ۸۲.۹٪ در دسته‌بندی نوع ملک و ۸۹٪ در شناسایی نوع آگهی‌دهنده

پایان
ممنون از توجه شما