1/5/2025

# Ensemble Learning

Ames Housing Dataset

Dr. F Keshavarze

Mohammad Mahdi Shafighi

SHAHED UNIVERSITY

فهرست مطالب


۱. توضیح داده‌ها و پیش‌پردازش‌های انجام‌شده : ..... ۳
۲. پیش‌پردازش داده‌ها شامل مراحل زیر بود: ..... ۳
۱. پاک‌سازی داده‌ها: ..... ۳
۲. تحلیل داده‌های اکتشافی:(EDA) ..... ۳
۳. مهندسی ویژگی‌ها: ..... ۳
٤. پیش‌پردازش نهایی داده‌ها: ..... ٤
۳. توضیح الگوریتم‌های پایه و روش‌هایEnsemble ..... ٤
●مدل‌های پایه: ..... ٤
●روش‌هایEnsemble : ..... ٤
٤. نتایج ارزیابی مدل‌ها ..... ٤
٥. نتایج ارزیابی مدل‌ها: ..... ٥
تحلیل نتایج ..... ٥
پیشنهادات برای بهبود بیشتر ..... ٦
I. تنظیم هایپرپارامترها: ..... ٦
II. افزایش ویژگی‌های مشتق‌شده: ..... ٦
III. استفاده ازEnsemble های پیچیده‌تر: ..... ٦
IV. بهبود داده‌ها: ..... ٦
V. بررسی حذف ویژگی‌های کم‌اهمیت: ..... ٦
جمع‌بندی: ..... ۷

بسم الله الرحمن الرحيم

# گزارش تحلیل و نتایج پروژه Ames Housing Data

## ۱. توضیح داده‌ها و پیش‌پردازش‌های انجام‌شده :

داده‌های استفاده‌شده شامل اطلاعات مرتبط با ویژگی‌های خانه‌ها و قیمت نهایی فروش آن‌ها

ستون SalePrice در منطقهAmes Iowa است. ویژگی‌ها به دو دسته کلی تقسیم می‌شوند:

ویژگی‌های عددی مانندGrLivArea ، LotArea، وYearBuilt

ویژگی‌های غیر عددی مانندNeighborhood ، BldgType، وHouseStyle

۲. پیش‌پردازش داده‌ها شامل مراحل زیر بود:

۱. پاکسازی داده‌ها:

- حذف مقادیر تکراری
- پر کردن مقادیر گمشده در ستون‌های عددی با مقدار میانه و در ستون‌های غیر عددی با پرتکرارترین مقدار
- حذف رکوردهایی که همچنان مقادیر گمشده داشتند.

۲. تحلیل داده‌های اکتشافی:(EDA)

- بررسی توزیع متغیر هدف (SalePrice) با استفاده از هیستوگرام.
- محاسبه ماتریس همبستگی ویژگی‌های عددی و نمایش آن با Heatmap
- ایجاد نمودار پراکندگی (scatter plot) برای بررسی رابطه GrLivArea و TotalBsmtSF با SalePrice

۳. مهندسی ویژگی‌ها:

- افزودن ویژگی جدید TotalLivingArea (مجموع مساحت زندگی در طبقات اصلی و زیرزمین)

### ٤. پیش‌پردازش نهایی داده‌ها:

- تقسیم داده‌ها به ویژگی‌ها (X) و متغیر هدف(y)
- دسته‌بندی ستون‌ها به ویژگی‌های عددی و غیر عددی
- استفاده از استاندارد سازی برای ستون‌های عددی و تبدیل OneHotEncoding برای ستون‌های غیر عددی

## ۳. توضیح الگوریتم‌های پایه و روش‌هایEnsemble

- مدل‌های پایه:

Linear Regression: یک مدل خطی ساده که برای پیش‌بینی متغیرهای پیوسته استفاده می‌شود.

Decision Tree: الگوریتمی مبتنی بر درخت که تصمیمات را بر اساس بیشترین کاهش خطا (مانندMSE ) اتخاذ می‌کند.

Random Forest: یک مدل Bagging که مجموعه‌ای از درخت‌های تصمیم‌گیری را ایجاد می‌کند و نتایج را ترکیب می‌کند.

- روش‌هایEnsemble :

Bagging (Random Forest): الگوریتمی که نمونه‌گیری Bootstrap انجام داده و درخت‌های تصمیم متعدد را ترکیب می‌کند.

Boosting (Gradient Boosting): الگوریتمی که مدل‌های متوالی ایجاد کرده و خطای نمونه‌های قبلی را بهبود می‌دهد.

Stacking: ترکیب چندین مدل پایه (مانند Random Forest وGradient Boosting ) و آموزش یک مدل متا برای بهبود عملکرد.

## ٤. نتایج ارزیابی مدل‌ها

در این پروژه، از معیارهای زیر برای ارزیابی عملکرد مدل‌ها استفاده شد:

Mean Squared Error (MSE): میانگین مربع خطاها.

Root Mean Squared Error (RMSE): ریشه دوم میانگین مربع خطاها.

$R^2$ Score: میزان واریانس توضیح داده‌شده توسط مدل. ( اضافی شده )

## ۵. نتایج ارزیابی مدل‌ها:

مدل‌هایBagging ، Boosting، و Stacking با استفاده از داده‌های Ames Housing مورد ارزیابی قرار گرفتندکه نتایج  از این قرار می باشد :

| $R^2$ Score | MSE | RMSE | MAE | مدل |
|---|---|---|---|---|
| 0.9130 | 697,214,755.73 | 26,404.82 | 15,823.80 | Bagging |
| 0.9119 | 706,663,100.46 | 26,583.14 | 15,249.18 | Boosting |
| 0.9075 | 741,483,154.40 | 27,230.19 | 17,135.91 | Stacking |

## تحلیل نتایج

۱. **Bagging (Random Forest)**

- بهترین عملکرد در معیارهای MAE و.RMSE
- این مدل توانایی بسیار خوبی در کاهش خطاهای پیش‌بینی دارد، به دلیل استفاده از چندین درخت تصمیم و ترکیب نتایج آن‌ها.

۲. **Boosting (Gradient Boosting)**

- عملکرد کمی ضعیف‌تر از Bagging در معیارهای RMSE و.$R^2$ Score
- مدل Boosting به دلیل افزایش تدریجی و بهبود خطاهای مدل‌های قبلی، مناسب پیش‌بینی‌های پیچیده‌تر است.

۳. **Stacking**

- عملکرد نسبتاً ضعیف‌تر در مقایسه با Bagging وBoosting

- این مدل به دلیل پیچیدگی بیشتر و وابستگی به کیفیت مدل‌های پایه، ممکن است در داده‌های خاص کارایی پایین‌تری داشته باشد.

## پیشنهادات برای بهبود بیشتر

I. تنظیم هایپرپارامترها:

- استفاده از GridSearchCV یا RandomizedSearchCV برای یافتن بهترین مقادیر هایپرپارامترهای مدل‌ها.
- برای مثال:
  - تعداد درخت‌ها (n_estimators) و عمق درخت‌ها (max_depth) در Random Forest.
  - نرخ یادگیری (learning_rate) و تعداد مراحل تقویت درGradient Boosting

II. افزایش ویژگی‌های مشتق‌شده:

- شناسایی و افزودن ویژگی‌های جدید مانند تعامل بین متغیرها یا استفاده از تحلیل‌های دامنه تخصصی (مانند نسبت مساحت زیرزمین به کل مساحت خانه).

III. استفاده ازEnsemble های پیچیده‌تر:

- ترکیب مدل‌های دیگری مانند XGBoost یا LightGBM با مدل‌های فعلی.

IV. بهبود داده‌ها:

- افزایش تعداد داده‌ها یا استفاده از تکنیک‌هایی مانند Data Augmentation برای کاهش خطاها.

V. بررسی حذف ویژگی‌های کم‌اهمیت:

- ویژگی‌هایی که همبستگی کمتری با متغیر هدف دارند، می‌توانند حذف شوند تا پیچیدگی مدل کاهش یابد.

## جمع‌بندی:

مدل Bagging با MAE کمترین و $R^2$ Score بالاترین، بهترین عملکرد را در این تحلیل ارائه داد. مدل‌های Boostingو Stacking نیز با وجود پیچیدگی بیشتر، عملکرد قابل‌قبولی داشتند. برای بهبود بیشتر، تنظیم دقیق هایپرپارامترها و استخراج ویژگی‌های جدید پیشنهاد می‌شود.