گزارش تمرین سوم درس داده کاوی

سید محمد امین حسینی (۹۹۴۲۲۰۵۷)

۱)

کرنل خطی. ساده‌ترین و پایه‌ترین نوع کرنل است که ماهیت آن معمولاً یک بعدی است. در صورت وجود تعداد بسیاری از ویژگی‌ها بهترین عملکرد را دارد. این کرنل بیشتر برای مسائل طبقه‌بندی متن استفاده می شود زیرا بیشتر این نوع مسائل را می توان به صورت خطی جدا کرد. همچنین این توابع سریعتر از توابع دیگر هستند.

کرنل چند جمله ای. این نمای جنرالایزتر شده ای از کرنل خطی است. به اندازه توابع دیگر مورد استفاده قرار نمی گیرد زیرا از کارآیی و دقت کمتری برخوردار است.

کرنل Gaussian Radial Basis (RBF). یکی از بهترین و مورد استفاده ترین توابع کرنل در svm است. معمولاً برای داده های غیر خطی انتخاب می شود. وقتی دانش قبلی از داده ها وجود نداشته باشد، به جداسازی مناسب کمک می کند.

کرنل سیگموئید. این تابع بیشتر برای شبکه های عصبی ترجیح داده می شود. این کرنل شبیه مدل پرسپترون دو لایه ای است که به عنوان یک تابع فعال ساز برای نورون ها عمل می کند.

کرنل ANOVA. معمولاً در مسائل رگرسیون چند بعدی عملکرد خوبی دارد.

۲)

در ادامه نتایج اعمال SVM با پارامترهای مختلف را بر روی دیتاست کلاس بندی دیتای موبایل مشاهده می کنیم.

۳)

تولرانس: 10^{-3} - کرنل: خطی

	precision	recall	f1-score	support
0	0.99	0.96	0.97	94
1	0.90	0.97	0.93	102
2	0.96	0.88	0.92	120
3	0.92	0.98	0.95	84
accuracy			0.94	400
macro avg	0.94	0.95	0.94	400
weighted avg	0.94	0.94	0.94	400

تولرانس: 10^{-3} - کرنل: RBF

	precision	recall	f1-score	support
0	0.96	0.95	0.95	94
1	0.84	0.84	0.84	102
2	0.84	0.84	0.84	120
3	0.91	0.92	0.91	84
accuracy			0.88	400
macro avg	0.89	0.89	0.89	400
weighted avg	0.88	0.88	0.88	400

تولرانس: 10^{-3} - کرنل: چندجمله ای

	precision	recall	f1-score	support
0	0.88	0.84	0.86	94
1	0.66	0.74	0.70	102
2	0.72	0.69	0.71	120
3	0.83	0.81	0.82	84
accuracy			0.76	400
macro avg	0.77	0.77	0.77	400
weighted avg	0.77	0.76	0.76	400

تولرانس: 10^{-3} - کرنل: سیگموئید

	precision	recall	f1-score	support
0	0.96	0.94	0.95	94
1	0.85	0.92	0.89	102
2	0.95	0.86	0.90	120
3	0.92	0.99	0.95	84
accuracy			0.92	400
macro avg	0.92	0.93	0.92	400
weighted avg	0.92	0.92	0.92	400

تولرانس: 10^{-1} - کرنل: خطی

	precision	recall	f1-score	support
0	0.99	0.97	0.98	94
1	0.91	0.98	0.94	102
2	0.97	0.88	0.93	120
3	0.92	0.98	0.95	84
accuracy			0.95	400
macro avg	0.95	0.95	0.95	400
weighted avg	0.95	0.95	0.95	400

تولرانس: 10^{-1} - کرنل: RBF

	precision	recall	f1-score	support
0	0.96	0.94	0.95	94
1	0.83	0.84	0.83	102
2	0.83	0.83	0.83	120
3	0.90	0.90	0.90	84
accuracy			0.88	400
macro avg	0.88	0.88	0.88	400
weighted avg	0.88	0.88	0.88	400

تولرانس: 10^{-1} - کرنل: چندجمله ای

	precision	recall	f1-score	support
0	0.88	0.84	0.86	94
1	0.66	0.74	0.70	102
2	0.72	0.69	0.71	120
3	0.83	0.81	0.82	84
accuracy			0.76	400
macro avg	0.77	0.77	0.77	400
weighted avg	0.77	0.76	0.76	400

تولرانس: 10^{-1} - کرنل: سیگموئید

	precision	recall	f1-score	support
0	0.96	0.94	0.95	94
1	0.85	0.91	0.88	102
2	0.94	0.85	0.89	120
3	0.91	0.99	0.95	84
accuracy			0.92	400
macro avg	0.92	0.92	0.92	400
weighted avg	0.92	0.92	0.91	400

با بررسی نتایج مدل‌ها مشاهده می شود که بر رو این دیتاست مدل SVM با کرنل خطی و مقدار تولرانس 10^{-1} بهترین نتیجه و SVM با کرنل چندچمله ای بدترین نتیجه را به همراه داشته است. بنابراین داده های این دیتاست به صورت خطی جداشونده هستند و نیازی به استفاده از کرنل های پیچیده تر نیست.

٤)

در SVM پکیج sklearn، پارامتر c تعیین کننده میزان مارجین است. به این صورت که هر چه به ۱ نزدیک تر باشد، مدل در کلاس بندی اشتباه بیشتر جریمه شده و مارجین باریک تر خواهد بود و هر چه c به ۰ نزدیک تر باشد برعکس. برای این مسئله چهار مقدار ۰/۲، ۰/٤، ۰/٦ و ۰/۸ برای c در نظر گرفته شده که با چهار کرنل تست شده است.

کرنل: خطی

Regularization parameter = 0.2

	precision	recall	f1-score	support
0	0.96	0.96	0.96	94
1	0.90	0.94	0.92	102
2	0.97	0.86	0.91	120
3	0.89	0.99	0.94	84
accuracy			0.93	400
macro avg	0.93	0.94	0.93	400
weighted avg	0.93	0.93	0.93	400

Regularization parameter = 0.6

	precision	recall	f1-score	support
0	0.98	0.96	0.97	94
1	0.89	0.96	0.92	102
2	0.96	0.87	0.91	120
3	0.91	0.98	0.94	84
accuracy			0.94	400
macro avg	0.94	0.94	0.94	400
weighted avg	0.94	0.94	0.93	400

Regularization parameter = 0.4

	precision	recall	f1-score	support
0	0.96	0.96	0.96	94
1	0.89	0.94	0.91	102
2	0.96	0.87	0.91	120
3	0.91	0.98	0.94	84
accuracy			0.93	400
macro avg	0.93	0.94	0.93	400
weighted avg	0.93	0.93	0.93	400

Regularization parameter = 0.8

	precision	recall	f1-score	support
0	0.99	0.96	0.97	94
1	0.89	0.97	0.93	102
2	0.96	0.88	0.92	120
3	0.92	0.98	0.95	84
accuracy			0.94	400
macro avg	0.94	0.94	0.94	400
weighted avg	0.94	0.94	0.94	400

کرنل: RBF

Regularization parameter = 0.6

	precision	recall	f1-score	support
0	0.94	0.90	0.92	94
1	0.79	0.86	0.82	102
2	0.85	0.82	0.83	120
3	0.92	0.90	0.91	84
accuracy			0.87	400
macro avg	0.87	0.87	0.87	400
weighted avg	0.87	0.87	0.87	400

Regularization parameter = 0.2

	precision	recall	f1-score	support
0	0.95	0.86	0.91	94
1	0.70	0.88	0.78	102
2	0.83	0.72	0.77	120
3	0.89	0.88	0.89	84
accuracy			0.83	400
macro avg	0.84	0.84	0.84	400
weighted avg	0.84	0.83	0.83	400

Regularization parameter = 0.8

	precision	recall	f1-score	support
0	0.95	0.93	0.94	94
1	0.80	0.85	0.82	102
2	0.84	0.81	0.83	120
3	0.90	0.90	0.90	84
accuracy			0.87	400
macro avg	0.87	0.87	0.87	400
weighted avg	0.87	0.87	0.87	400

Regularization parameter = 0.4

	precision	recall	f1-score	support
0	0.95	0.88	0.92	94
1	0.75	0.87	0.81	102
2	0.81	0.79	0.80	120
3	0.91	0.85	0.88	84
accuracy			0.84	400
macro avg	0.86	0.85	0.85	400
weighted avg	0.85	0.84	0.85	400

کرنل: چندجمله ای

Regularization parameter = 0.2

	precision	recall	f1-score	support
0	0.87	0.76	0.81	94
1	0.49	0.85	0.62	102
2	0.68	0.35	0.46	120
3	0.85	0.80	0.82	84
accuracy			0.67	400
macro avg	0.72	0.69	0.68	400
weighted avg	0.71	0.67	0.66	400

Regularization parameter = 0.6

	precision	recall	f1-score	support
0	0.87	0.80	0.83	94
1	0.65	0.74	0.69	102
2	0.71	0.71	0.71	120
3	0.84	0.79	0.81	84
accuracy			0.75	400
macro avg	0.77	0.76	0.76	400
weighted avg	0.76	0.75	0.75	400

Regularization parameter = 0.4

	precision	recall	f1-score	support
0	0.90	0.78	0.83	94
1	0.62	0.79	0.70	102
2	0.73	0.71	0.72	120
3	0.90	0.79	0.84	84
accuracy			0.76	400
macro avg	0.79	0.77	0.77	400
weighted avg	0.78	0.76	0.77	400

Regularization parameter = 0.8

	precision	recall	f1-score	support
0	0.88	0.83	0.85	94
1	0.65	0.73	0.69	102
2	0.72	0.68	0.70	120
3	0.82	0.82	0.82	84
accuracy			0.76	400
macro avg	0.77	0.77	0.77	400
weighted avg	0.76	0.76	0.76	400

کرنل: سیگموئید

Regularization parameter = 0.6

	precision	recall	f1-score	support
0	0.95	0.94	0.94	94
1	0.84	0.89	0.87	102
2	0.92	0.85	0.88	120
3	0.92	0.96	0.94	84
accuracy			0.91	400
macro avg	0.91	0.91	0.91	400
weighted avg	0.91	0.91	0.90	400

Regularization parameter = 0.2

	precision	recall	f1-score	support
0	0.99	0.95	0.97	94
1	0.84	0.95	0.89	102
2	0.93	0.82	0.87	120
3	0.91	0.96	0.94	84
accuracy			0.91	400
macro avg	0.92	0.92	0.92	400
weighted avg	0.92	0.91	0.91	400

Regularization parameter = 0.8

	precision	recall	f1-score	support
0	0.96	0.93	0.94	94
1	0.83	0.93	0.88	102
2	0.96	0.84	0.90	120
3	0.92	0.99	0.95	84
accuracy			0.92	400
macro avg	0.92	0.92	0.92	400
weighted avg	0.92	0.92	0.92	400

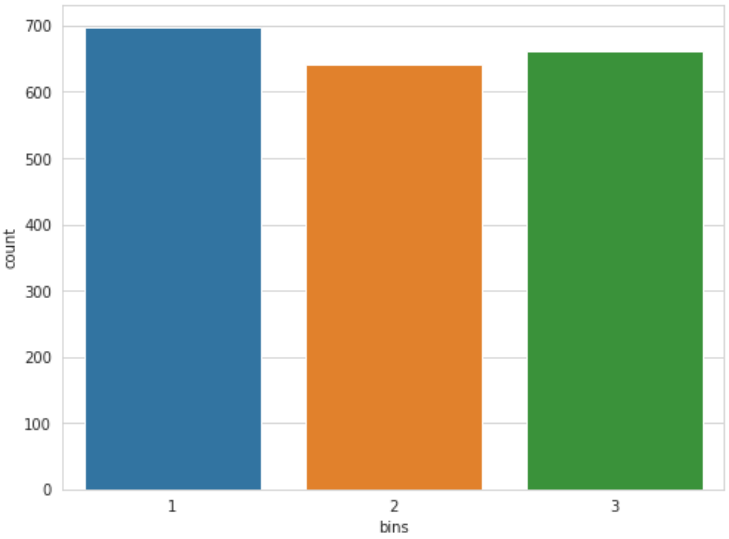
Regularization parameter = 0.4

	precision	recall	f1-score	support
0	0.95	0.94	0.94	94
1	0.85	0.92	0.89	102
2	0.96	0.85	0.90	120
3	0.91	0.99	0.95	84
accuracy			0.92	400
macro avg	0.92	0.92	0.92	400
weighted avg	0.92	0.92	0.92	400

۵)

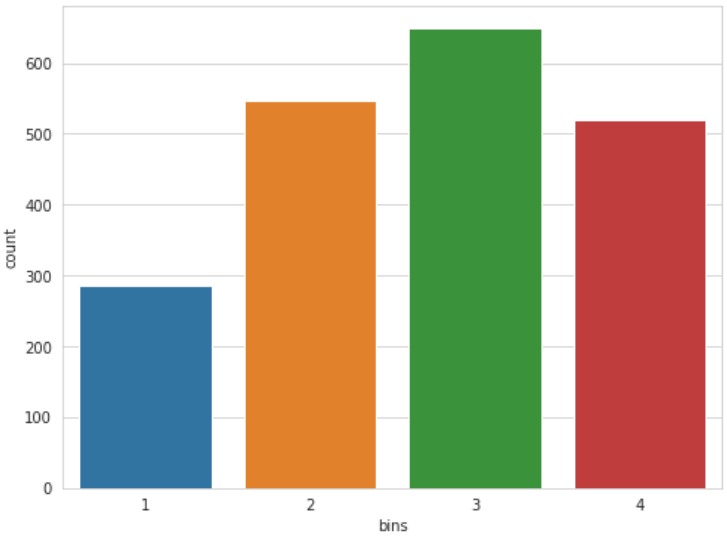
Binning:

سه bin به روش min و max با فواصل یکسان



	precision	recall	f1-score	support
0	0.96	0.97	0.96	94
1	0.87	0.93	0.90	102
2	0.95	0.86	0.90	120
3	0.93	0.98	0.95	84
accuracy			0.93	400
macro avg	0.93	0.93	0.93	400
weighted avg	0.93	0.93	0.93	400

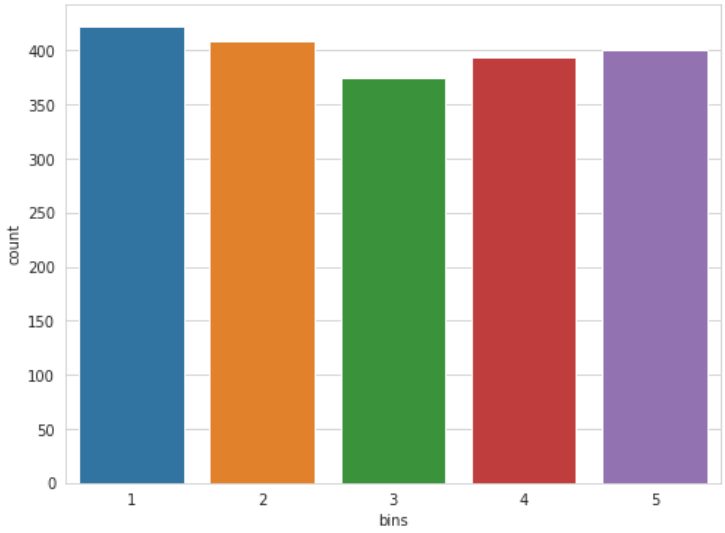
چهار bin به روش min و max با فواصل غیر یکسان و انتخاب دستی



	precision	recall	f1-score	support
0	0.95	0.94	0.94	94
1	0.88	0.92	0.90	102
2	0.95	0.89	0.92	120
3	0.93	0.96	0.95	84
accuracy			0.93	400
macro avg	0.93	0.93	0.93	400
weighted avg	0.93	0.93	0.93	400

پنج bin به روش min و max با فواصل یکسان

	precision	recall	f1-score	support
0	0.96	0.97	0.96	94
1	0.90	0.92	0.91	102
2	0.95	0.88	0.91	120
3	0.93	0.98	0.95	84
accuracy			0.93	400
macro avg	0.93	0.94	0.93	400
weighted avg	0.93	0.93	0.93	400



One Hot Encoding:

اعمال OHE بر فیچرهای fc، m_dep، n_cores، pc

	precision	recall	f1-score	support
0	0.99	0.98	0.98	94
1	0.89	0.95	0.92	102
2	0.94	0.85	0.89	120
3	0.91	0.96	0.94	84
accuracy			0.93	400
macro avg	0.93	0.94	0.93	400
weighted avg	0.93	0.93	0.93	400

log transform:

	precision	recall	f1-score	support
0	0.99	0.97	0.98	94
1	0.91	0.98	0.94	102
2	0.97	0.88	0.93	120
3	0.92	0.98	0.95	84
accuracy			0.95	400
macro avg	0.95	0.95	0.95	400
weighted avg	0.95	0.95	0.95	400

تبدیلات بر روی داده ها باعث بهبود سازماندهی آنها می شود. استفاده و کار با داده های تبدیل شده ممکن است برای انسان و کامپیوتر آسان تر باشد. این نوع تبدیلات معمولا کیفیت داده ها را بهبود می بخشد و سازگاری بین کاربردها،

سیستم ها و انواع داده ها را تسهیل می کند. ممکن است لازم باشد داده های مورد استفاده برای اهداف مختلف به روش های متفاوتی تبدیل شوند. از روش های مختلف تبدیلات می توان به تبدیل با توان، استانداردسازی، نرمال کردن با میانگین و مقیاس بندی با min و max اشاره کرد. در خصوص تبدیل log، به طوری کلی این تبدیل کجی داده های اصلی را کاهش داده یا از بین می برد.

همان طور که مشاهده می شود تبدیل log بر روی این دیتاست بهبود خیلی جزیی ۱ تا ۲ درصد به همراه داشته است.

ایجاد فیچر جدید area:

در این بخش با ضرب مقادیر فیچر px_height در px_width ویژگی مساحت به دیتاست اضافه شده است.

	precision	recall	f1-score	support
0	0.97	0.97	0.97	94
1	0.91	0.95	0.93	102
2	0.96	0.88	0.92	120
3	0.92	0.98	0.95	84
accuracy			0.94	400
macro avg	0.94	0.94	0.94	400
weighted avg	0.94	0.94	0.94	400

٦)

حال یک مدل SVM بر روی دیتاست با تمام تغییرات بالا (از binning نوع سوم استفاده شده است) می سازیم. (نتایج مدل های جداگانه در سوال قبل در ادامه هر تغییر آورده شده است.)

	precision	recall	f1-score	support
0	0.97	0.98	0.97	94
1	0.88	0.93	0.90	102
2	0.92	0.81	0.86	120
3	0.87	0.95	0.91	84
accuracy			0.91	400
macro avg	0.91	0.92	0.91	400
weighted avg	0.91	0.91	0.91	400

هر کدام از این تغییرات به طور جداگانه یا تغییری در نتایج نکرده یا باعث بهبود جزیی شده اند. در حالی که اعمال همه تغییرات باهم کاهش دقت به همراه داشته است.

۷)

الگوریتم ID۳ اولین الگوریتم ابداع شده برای درخت های تصمیم است. این الگوریتم در هر مرحله با تکرار ویژگی ها را به دو یا چند گروه تقسیم می کند. ID۳ برای ایجاد درخت تصمیم از روش حریصانه از بالا به پایین استفاده می کند. رویکرد بالا به پایین به این معنی که از بالا شروع به ساختن درخت کرده و رویکرد حریصانه به این معنی که در هر تکرار بهترین ویژگی در لحظه فعلی را برای ایجاد گره انتخاب می کنیم.

C٤/٥ جانشین ID۳ است و با تعریف پویا یک ویژگی گسسته (بر اساس متغیرهای عددی) که مقدار ویژگی پیوسته را به یک مجموعه گسسته از فواصل تقسیم می کند، محدودیت لزوم کتگوریکال بودن ویژگی ها را از بین برد. C٤/٥ درختان آموزش دیده (یعنی خروجی الگوریتم ID۳) را به مجموعه ای از قوانین if-then تبدیل می کند. دقت هر قانون سپس ارزیابی می شود تا ترتیب استفاده از آنها تعیین شود.

C٥/٠ آخرین نسخه از ID۳ است. این روش از حافظه کمتری استفاده و قوانین کوچکتری نسبت به C٤/٥ ایجاد می کند، در حالی که دقت C٤/٥ بیشتر است.

CART یا درخت کلاس بندی و رگرسیون اغلب به عنوان اصطلاح درخت تصمیم استفاده می شود، گرچه ظاهراً معنای خاص تری دارد. به طور خلاصه، CART بسیار شبیه C٤/٥ است. یک تفاوت قابل توجه این است که CART درخت را بر اساس معیار تقسیم عددی که بطور بازگشتی بر روی داده اعمال می شود می سازد، در حالی که C٤/٥ شامل مرحله میانیِ مجموعه قانون های ساخت است.

۸)

	precision	recall	f1-score	support
0	0.91	0.85	0.88	94
1	0.73	0.83	0.78	102
2	0.86	0.74	0.79	120
3	0.86	0.93	0.89	84
accuracy			0.83	400
macro avg	0.84	0.84	0.84	400
weighted avg	0.84	0.83	0.83	400

(۹

۲ :min_samples_split ،None :max_depth ،'best' :splitter ،'entropy' :criterion

	precision	recall	f1-score	support
0	0.92	0.91	0.92	94
1	0.82	0.82	0.82	102
2	0.85	0.83	0.84	120
3	0.89	0.93	0.91	84
accuracy			0.87	400
macro avg	0.87	0.88	0.87	400
weighted avg	0.87	0.87	0.87	400

۲ :min_samples_split ،None :max_depth ،'random' :splitter ،'gini' :criterion

	precision	recall	f1-score	support
0	0.91	0.87	0.89	94
1	0.69	0.76	0.73	102
2	0.75	0.66	0.70	120
3	0.80	0.88	0.84	84
accuracy			0.78	400
macro avg	0.79	0.79	0.79	400
weighted avg	0.78	0.78	0.78	400

۲ :min_samples_split ،۳ :max_depth ،'best' :splitter ،'gini' :criterion

	precision	recall	f1-score	support
0	0.90	0.76	0.82	94
1	0.63	0.79	0.70	102
2	0.74	0.63	0.68	120
3	0.78	0.83	0.80	84
accuracy			0.74	400
macro avg	0.76	0.75	0.75	400
weighted avg	0.76	0.74	0.75	400

۵۰۰ :min_samples_split ،None :max_depth ،'best' :splitter ،'gini' :criterion

	precision	recall	f1-score	support
0	0.87	0.78	0.82	94
1	0.63	0.76	0.69	102
2	0.74	0.63	0.68	120
3	0.78	0.83	0.80	84
accuracy			0.74	400
macro avg	0.75	0.75	0.75	400
weighted avg	0.75	0.74	0.74	400

criterion: 'gini'، splitter: 'best'، max_depth: ۲، min_samples_split: ۳۰۰

	precision	recall	f1-score	support
0	0.87	0.78	0.82	94
1	0.63	0.76	0.69	102
2	0.74	0.63	0.68	120
3	0.78	0.83	0.80	84
accuracy			0.74	400
macro avg	0.75	0.75	0.75	400
weighted avg	0.75	0.74	0.74	400

criterion: 'entropy'، splitter: 'best'، max_depth: ٤، min_samples_split: ۱۰

	precision	recall	f1-score	support
0	0.79	0.94	0.85	94
1	0.73	0.54	0.62	102
2	0.70	0.76	0.73	120
3	0.82	0.81	0.81	84
accuracy			0.76	400
macro avg	0.76	0.76	0.75	400
weighted avg	0.75	0.76	0.75	400

در ابتدا تغییر criterion از gini به entropy باعث ٤ درصد بهبود دقت شده است. پارامتر splitter نیز با تغییر از best به random کاهش ۵ درصدی دقت را به همراه داشته است. همچنین نسبت دادن مقادیر کم به max_depth و مقادیر زیاد به min_samples_split عملکرد مدل را تحت تاثیر منفی قرار می دهد.

۱۰)

در روش های هرس با برداشتن قسمت هایی از درخت که قدرت طبقه بندی نمونه ها را ندارند، اندازه درخت تصمیم کاهش می یابد. درخت های تصمیم نسبت به دیگر الگوریتم های یادگیری ماشین بیشتر مستعد بیش برازش (overfitting) هستند و هرس موثر می تواند این احتمال را کاهش دهد. در ادامه برخی از رویکردهای هرس به طور خلاصه شرح داده شده است:

Reduced Error Pruning (REP): این روش از نظر مفهومی ساده ترین روش است و از مجموعه هرس برای ارزیابی کارایی یک زیرشاخه Tmax استفاده می کند. فرآیند هرس با درخت کامل Tmax شروع می شود و برای هر گره داخلی Tmax، تعداد خطاهای کلاس بندی صورت گرفته در مجموعه هرس هنگام نگهداری زیرشاخه Tt را با تعداد خطاهای کلاس بندی هنگام تبدیل t، که مرتبط با بهترین کلاس است، به برگ مقایسه می کند.

Pessimistic Error Pruning (PEP): این روش هرس مانند روش قبلی، برای پرورش و هرس درخت از یک مجموعه آموزش یکسان استفاده می شود. بدیهی است که میزان خطای آشکار، یعنی میزان خطا در مجموعه آموزش، اریب است و

نمی توان از آن برای انتخاب بهترین درخت هرس شده استفاده کرد. به همین دلیل، در این روش یک پیوستگی برای توزیع دوجمله ای ارائه شده است که می تواند یک نرخ خطای واقعی تر را ارائه دهد.

Minimum Error Pruning (MEP): یک روش از پایین به بالا که به دنبال ایجاد یک درخت است که "میزان خطای مورد انتظار در یک مجموعه داده مستقل" را به حداقل برساند. این بدان معنا نیست که از یک مجموعه هرس استفاده می شود، بلکه صرفاً برای تخمین میزان خطا در نمونه های گم شده است.

Critical Value Pruning (CVP): این روش post-pruning شباهت زیادی به روش pre-pruningدارد. در واقع ، یک آستانه به نام مقدار بحرانی ، برای معیار انتخاب گره تنظیم شده است. سپس اگر مقدار بدست آمده با معیار انتخاب برای هر تست مرتبط با یال هایی که از آن گره خارج می شوند از مقدار بحرانی بیشتر نشود، گره داخلی درخت هرس می شود.

Cost-Complexity Pruning (CCP): این روش به الگوریتم هرس CART نیز معروف است که شامل دو مرحله است:

۱- انتخاب یک خانواده پارامتریک از زیرشاخه های Tmax، {T۰، T۱، ...، TL} ، با توجه به روش های ابتکاری.

۲- انتخاب بهترین درخت Ti با توجه به تخمین میزان خطای واقعی درختان در خانواده پارامتریک.

Error-Based Pruning (EBP): این روش هرس پیاده سازی شده در C٤/٥ است. این روش یک توسعه از روش PEP در نظر گرفته می شود، زیرا بر اساس یک تخمین بدبینانه تر از میزان خطای مورد انتظار است. هر دو روش از اطلاعات موجود در مجموعه آموزش ساخت و ساده سازی درختان استفاده می کنند.

۱۱)

معیارهای ارزیابی به ترتیب برای داده آموزش و تست، بدون به کارگیری هرس (بیش برازش شده)

	precision	recall	f1-score	support
0	1.00	1.00	1.00	406
1	1.00	1.00	1.00	398
2	1.00	1.00	1.00	380
3	1.00	1.00	1.00	416
accuracy			1.00	1600
macro avg	1.00	1.00	1.00	1600
weighted avg	1.00	1.00	1.00	1600

	precision	recall	f1-score	support
0	0.91	0.85	0.88	94
1	0.73	0.83	0.78	102
2	0.86	0.74	0.79	120
3	0.86	0.93	0.89	84
accuracy			0.83	400
macro avg	0.84	0.84	0.84	400
weighted avg	0.84	0.83	0.83	400

معیارهای ارزیابی به ترتیب برای داده آموزش و تست، با به کارگیری هرس (جلوگیری از بیش برازش)

	precision	recall	f1-score	support
0	0.96	0.94	0.95	406
1	0.87	0.93	0.90	398
2	0.92	0.88	0.90	380
3	0.96	0.96	0.96	416
accuracy			0.93	1600
macro avg	0.93	0.93	0.93	1600
weighted avg	0.93	0.93	0.93	1600

	precision	recall	f1-score	support
0	0.90	0.85	0.87	94
1	0.72	0.83	0.77	102
2	0.85	0.75	0.80	120
3	0.87	0.90	0.89	84
accuracy			0.83	400
macro avg	0.84	0.83	0.83	400
weighted avg	0.83	0.83	0.83	400

همانطور که مشاهده می شود، استفاده از روش هرس مانع رخ دادن بیش برازش در مدل شده است.

۱۲)

نتایج اجرای Random Forest

	precision	recall	f1-score	support
0	0.93	0.93	0.93	94
1	0.82	0.87	0.85	102
2	0.90	0.85	0.88	120
3	0.93	0.94	0.93	84
accuracy			0.89	400
macro avg	0.90	0.90	0.90	400
weighted avg	0.89	0.89	0.89	400

نتایج اجرای درخت تصمیم

	precision	recall	f1-score	support
0	0.91	0.85	0.88	94
1	0.73	0.83	0.78	102
2	0.86	0.74	0.79	120
3	0.86	0.93	0.89	84
accuracy			0.83	400
macro avg	0.84	0.84	0.84	400
weighted avg	0.84	0.83	0.83	400

با توجه به معیارهای ارزیابی، به کار گرفتن Random Forest برای این دیتاست باعث بهبود ٦ درصدی نتایج شده است. این بهبود در دقت به دلیل ذات این روش است که از ویژگی ensemble بهره می برد. Random Forest شامل چندین درخت تصمیم است که هر کدام یک پیش بینی برای داده ها ارائه می دهند. در نهایت درخت با بهترین دقت به عنوان پیش بینی کننده اصلی مدل انتخاب می شود.

(۱۳

یک مزیت قابل توجه درخت تصمیم این است که تمام نتایج احتمالی و مسیرهای منتهی به هدف را به نتیجه می رساند. این کار یک تجزیه و تحلیل جامع از پیامدهای هر شاخه ایجاد می کند و گره های تصمیم را که نیاز به تجزیه و تحلیل بیشتر دارند، شناسایی می کند.

در مقایسه با الگوریتم های دیگر، درختان تصمیم به تلاش کمتری برای آماده سازی داده ها در مرحله پیش پردازش نیاز دارند.

خواندن و تفسیر خروجی آنها آسان است، حتی بدون نیاز به دانش آماری. این به این دلیل است که نمایشی گرافیکی از مسئله و انواع جایگزین ها را در قالبی شهودی و ساده ارائه می دهد که نیازی به مصورسازی ندارد.

برخلاف سایر ابزارهای تصمیم گیری که به داده های کمّی جامع نیاز دارند، درختان تصمیم گیری برای حل مسائل ترکیبی از ویژگی های کتگوریکال، ویژگی های با مقدار واقعی و یا بعضاً همراه با نمونه های گم شده، انعطاف پذیر باقی می مانند.

(۱٤

Rule Induction یکی از مهمترین تکنیک های یادگیری ماشین است. از آنجا که قواعد پنهان در داده ها غالباً به شکل قوانین بیان می شوند، Rule Induction یکی از ابزارهای اساسی داده کاوی است. معمولاً این قوانین به صورت زیر هستند:

$$if\ (attribute - ۱;\ value - ۱) and\ (attribute - ۲;\ value - ۲) and\ \dots$$
$$and\ (attribute - n;\ value - n)\ then\ (decision;\ value)$$

برخی از این سیستم ها قوانین پیچیده تری را استخراج می کنند که در آنها مقادیر ویژگی ها ممکن است با تفریق برخی از مقادیر یا با یک زیرمجموعه مقدار از دامنه ویژگی بیان شوند.

یکی از کارآمد و مورد استفاده ترین این الگوریتم ها RIPPER (هرس افزایشی مکرر برای تولید کاهش خطا) است. این الگوریتم یک استراتژی تقسیم و غلبه را برای این کار اجرا می کند. Ripper اصطلاحاً هرس خطای کاهش یافته افزایشی (IREP) را برای استخراج مجموعه ای از قوانین اولیه برای هر کلاس اعمال می کند. سپس ، یک مرحله بهینه سازی اضافی هر قانون را در مجموعه فعلی به نوبت در نظر می گیرد و دو قانون جایگزین از آنها ایجاد می کند: یک قانون جایگزینی و یک قانون تجدید نظر. پس از آن بر اساس معیار حداقل طول توصیف در مورد اینکه مدل باید کدام یک از قانون اصلی، جایگزینی یا قانون تجدید نظر را نگه دارد، تصمیم گیری می شود.

PART یک الگوریتم درخت تصمیم جزئی است. این الگوریتم براساس استراتژی تقسیم و غلبه، مجموعه ای از قوانین را ایجاد می کند، تمام موارد را از مجموعه آموزش که تحت این قانون قرار دارند حذف می کند و به صورت بازگشتی ادامه می دهد تا زمانی که هیچ موردی باقی بماند. برای ایجاد یک قانون واحد، PART یک درخت تصمیم C٤٫٥ جزئی برای مجموعه موارد فعلی ایجاد می کند و برگ با بیشترین پوشش را به عنوان قانون جدید انتخاب می کند. سپس درخت تصمیم گیری

جزئی به همراه نمونه های تحت پوشش قانون جدید از داده های آموزش حذف می شود تا از تعمیم زودهنگام جلوگیری شود. این روند تکرار می شود تا زمانی که همه موارد با قوانین استخراج شده پوشش داده شوند.

FURIA نسخه بهبود یافته الگوریتم RIPPER است. FURIA از الگوریتم اصلاح شده RIPPER به عنوان مبنا استفاده می کند و قوانین مبهم و مجموعه قوانین غیر مرتب را می آموزد. نقطه قوت اصلی این الگوریتم روش کشش قانون است، که مشکل فشرده سازی نمونه های جدید را که در صورت طبقه بندی می توانند خارج از فضای تحت پوشش قوانین قبلی باشند، حل می کند. نمایش قوانین مبهم نیز پیشرفته است ، اساساً یک قانون مبهم از طریق جایگزینی فواصل با فواصل مبهم، یعنی مجموعه های مبهم با تابع عضویت ذوزنقه ای به دست می آید.

۱۵)

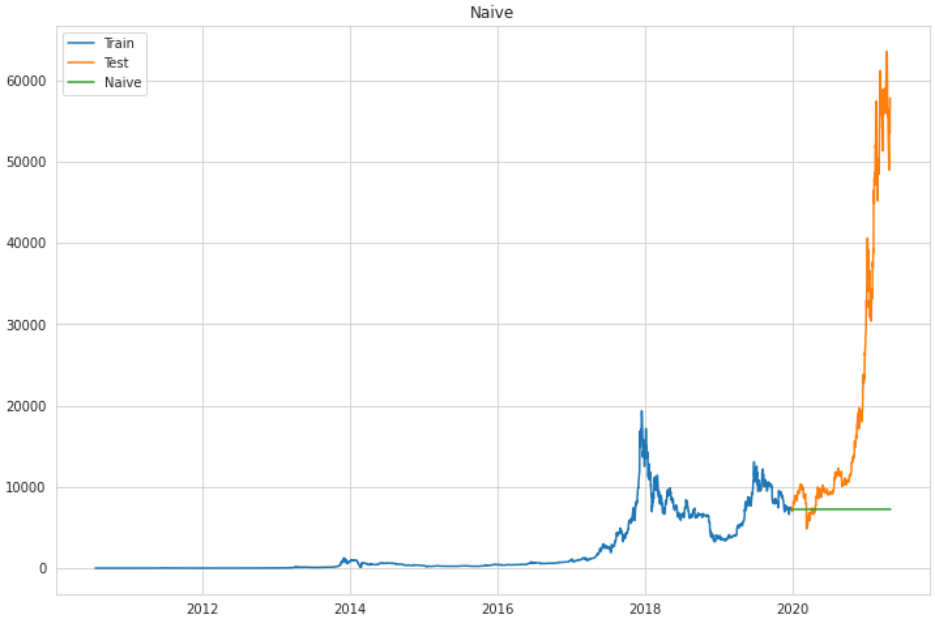
به طور کلی درخت های تصمیم توانایی برون یابی (extrapolation) مناسبی ندارند و از این جهت مدل کارایی برای پیش بینی و حل مسائل سری های زمانی نیستند. مانند رگرسیون خطی یا شبکه عصبی قابلیت گسترش پیش بینی تا بی نهایت را ندارند و فقط می تواند در حدود آنچه که قبلاً دیده و دریافت کرده پیش بینی کنند.

۱۶)

Date	Price	Open	High	Low	Vol.	Change %
2010-07-18	0.1	0.0	0.1	0.1	80.0	0.00
2010-07-19	0.1	0.1	0.1	0.1	570.0	0.00
2010-07-20	0.1	0.1	0.1	0.1	260.0	0.00
2010-07-21	0.1	0.1	0.1	0.1	580.0	0.00
2010-07-22	0.1	0.1	0.1	0.1	2160.0	0.00
...	...	...	...	...	...	...
2021-04-27	55036.5	54011.1	55427.8	53345.0	84080.0	1.88
2021-04-28	54841.4	55036.0	56419.9	53876.4	86960.0	-0.35
2021-04-29	53560.8	54838.6	55173.7	52400.0	83900.0	-2.34
2021-04-30	57720.3	53562.3	57925.6	53088.7	103740.0	7.77
2021-05-01	57807.1	57719.1	58449.4	57029.5	63410.0	0.15

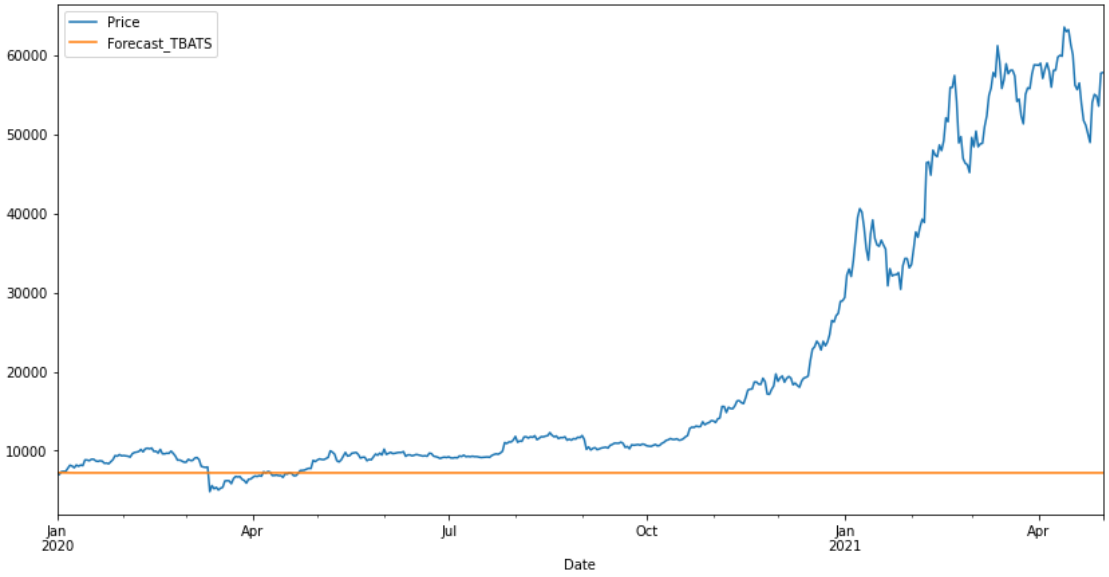
۱۷)

مدل: Naïve

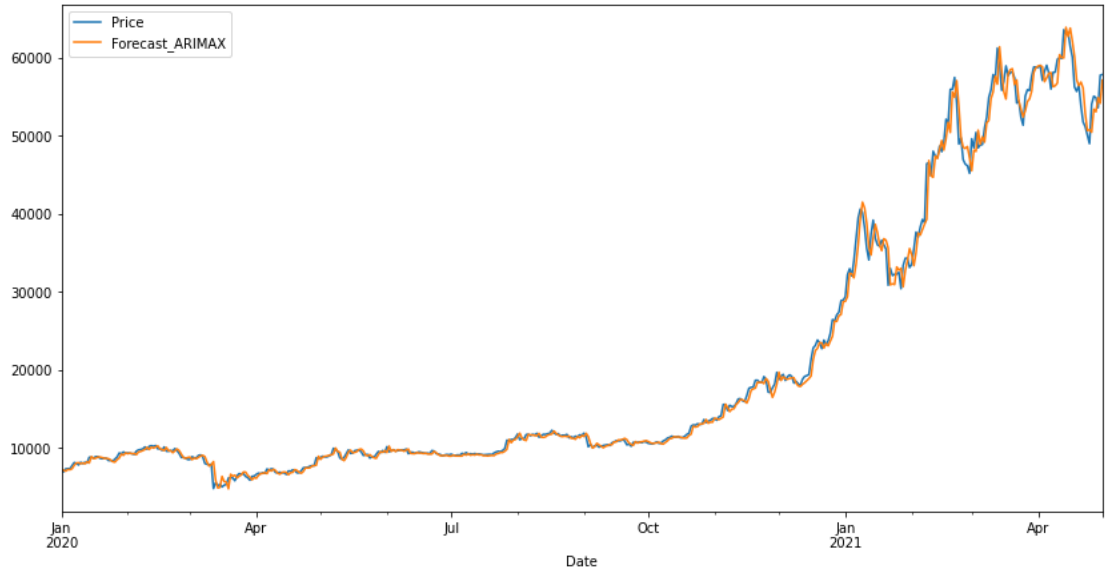


مدل: Seasonal Naïve

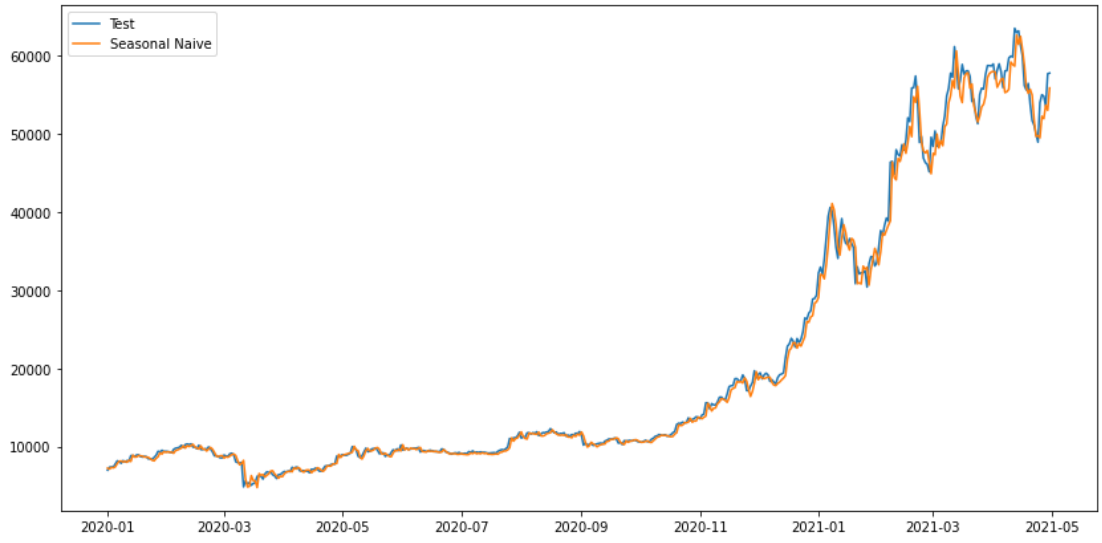
مدل: TBATS



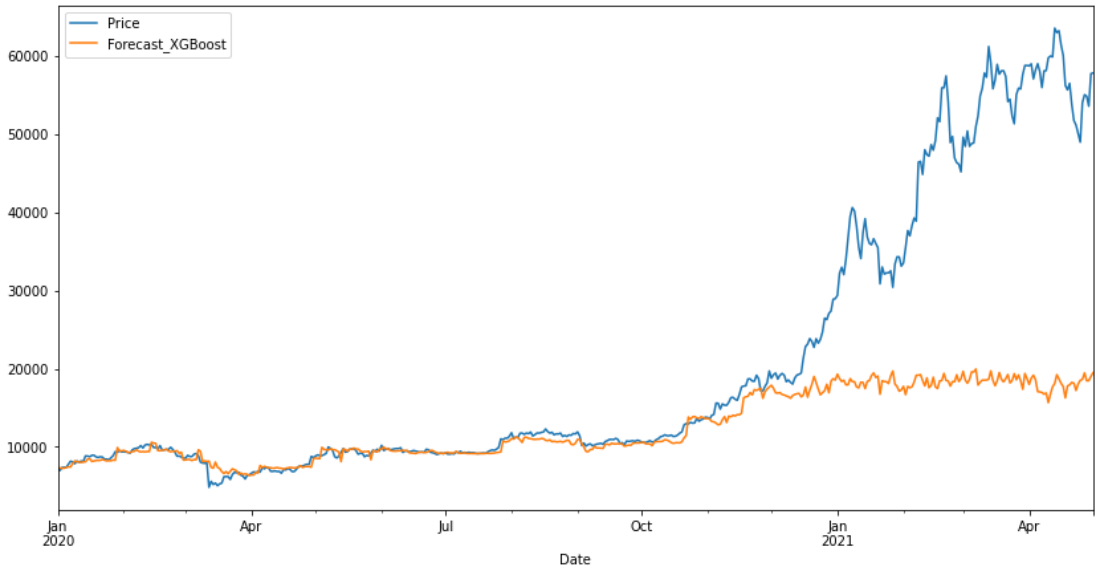
مدل: ARIMAX



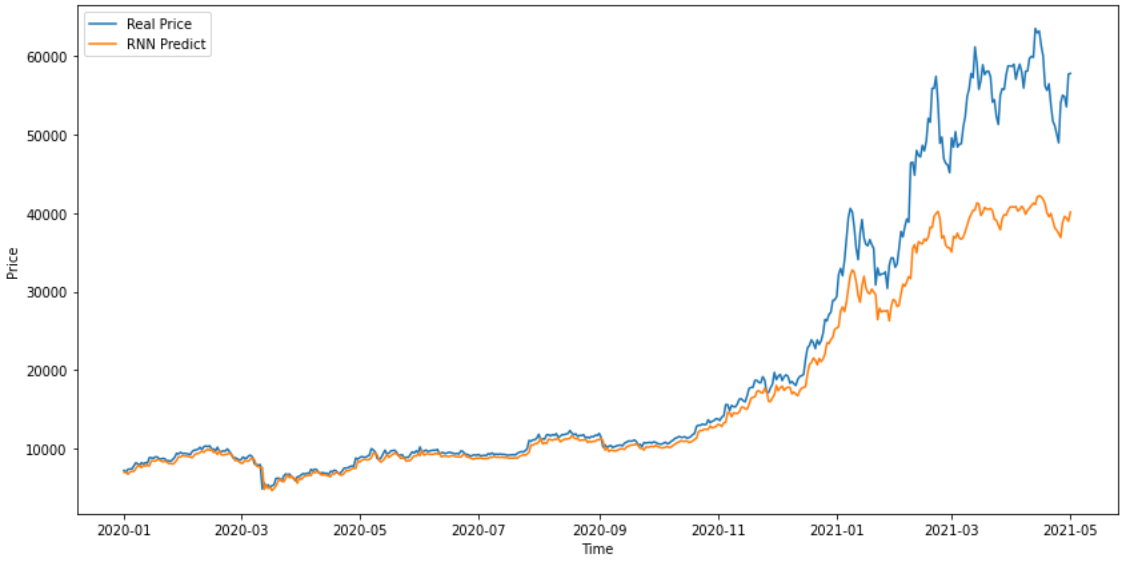
مدل: FaceBook Prophet



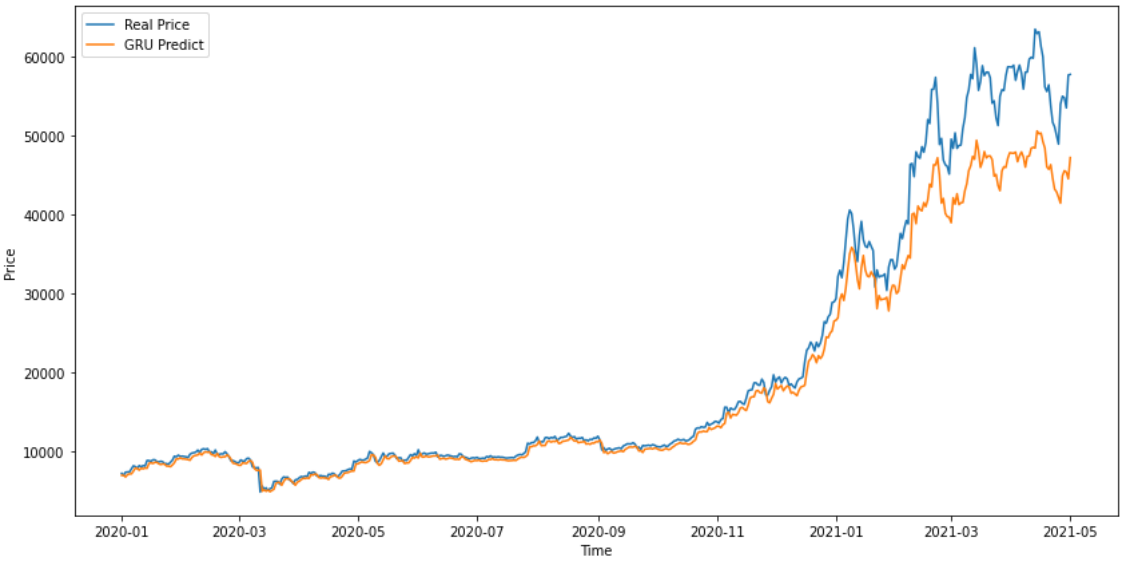
مدل: XG Boost



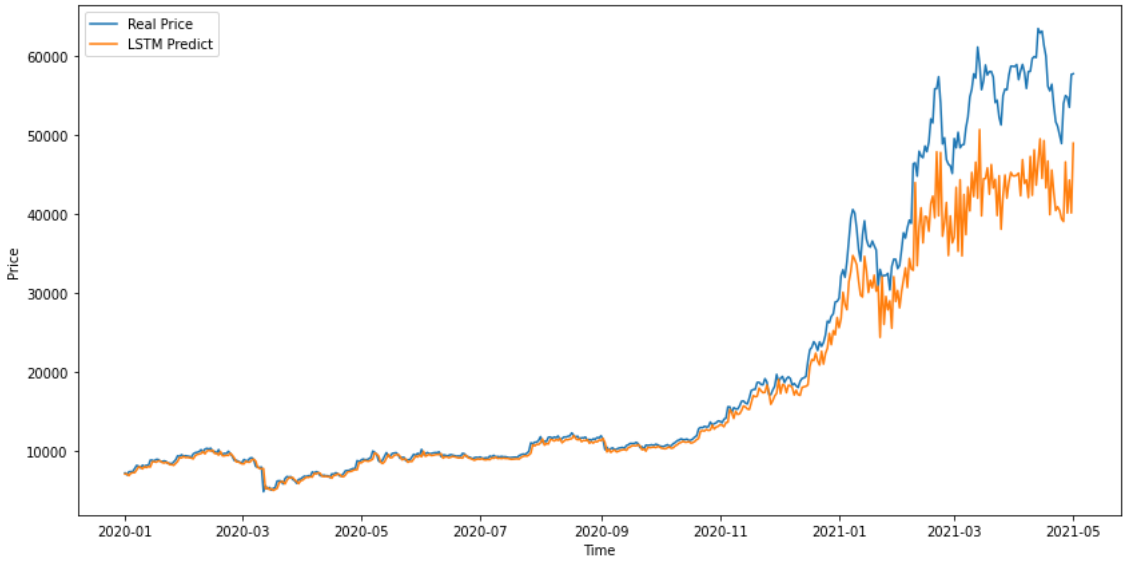
مدل: RNN



مدل: GRU



مدل: LSTM



خطا و دقت مدل ها:

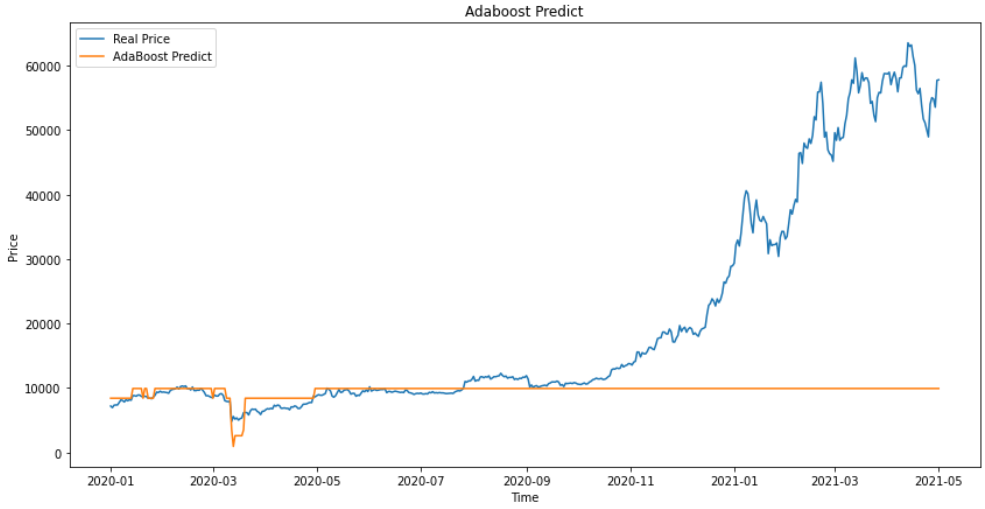
مدل	دقت (%)	خطا (RMSE)
Naïve	٤٫٣١	٢١٦٥٩٫١٣
Seasonal Naïve	١٫٠٢	٢١١٤٠٫٩٩
TBATS	٤٫٣١	٢١٦٦٠٫٠٨
ARIMAX	٨٢٫٥٤	١١٥٧٫٣٣
FB Prophet	٨١٫٦٨	١٢٣٨٫٤٤
XG Boost	٤٠٫٤٥	١٥٨٠٨٫٦٧
RNN	٣٥٫٩٣	٦٧٦١٫٧٨
GRU	٤٧٫٦٣	٤٢٢٤٫١٣
LSTM	٥٧٫٤٩	٥٣٥١٫٥٢

۱۸)

۱۹)

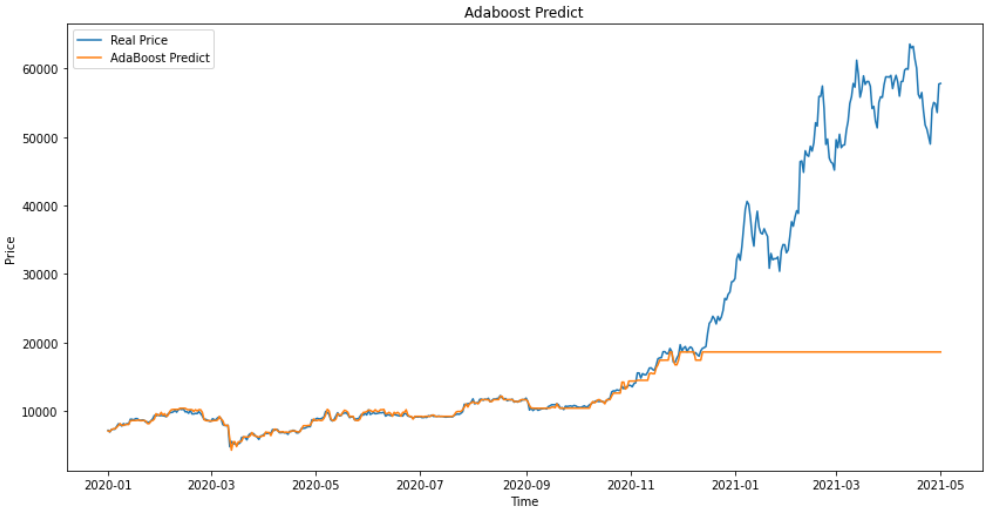
اجرای الگوریتم AdaBoost با پارامترهای مختلف:

max_depth: ۱، n_estimators: ۵۰، learning_rate: ۱



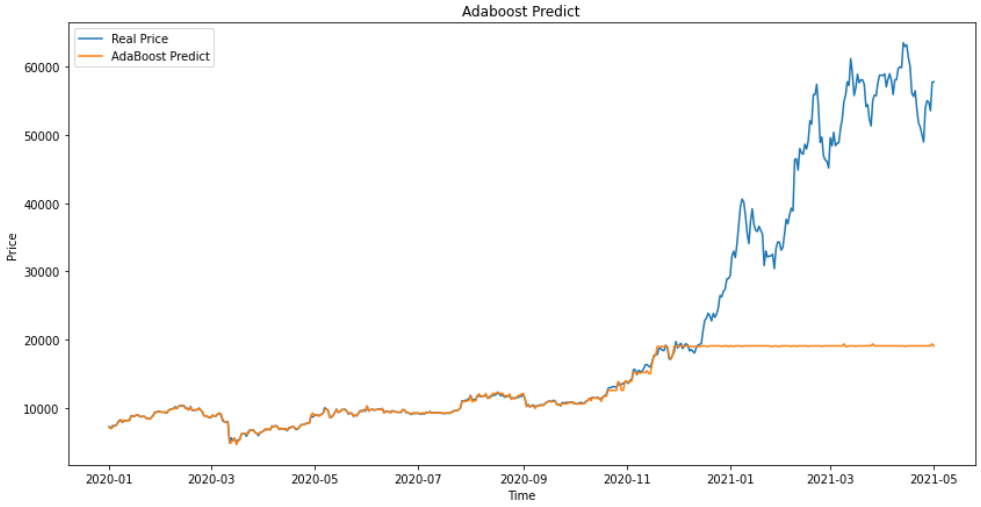
دقت: ۱۵۳۷٤/۵۵ خطا: ٦۸/۵۸

max_depth: ۳، n_estimators: ۵۰، learning_rate: ۱



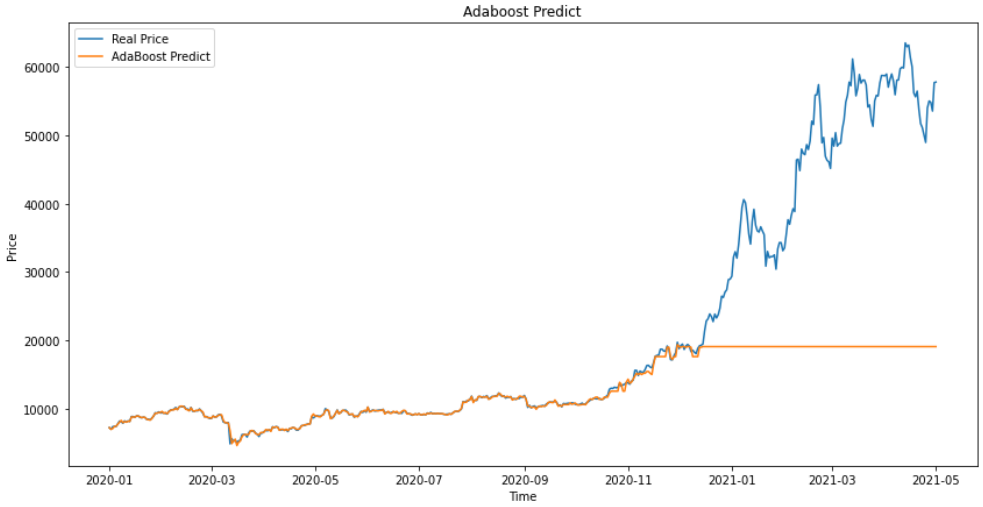
دقت: ۱۵۵۷۸/۰۷ خطا: ٦٤/۸۸

max_depth: ۱۰، n_estimators: ۱۰۰، learning_rate: ۱



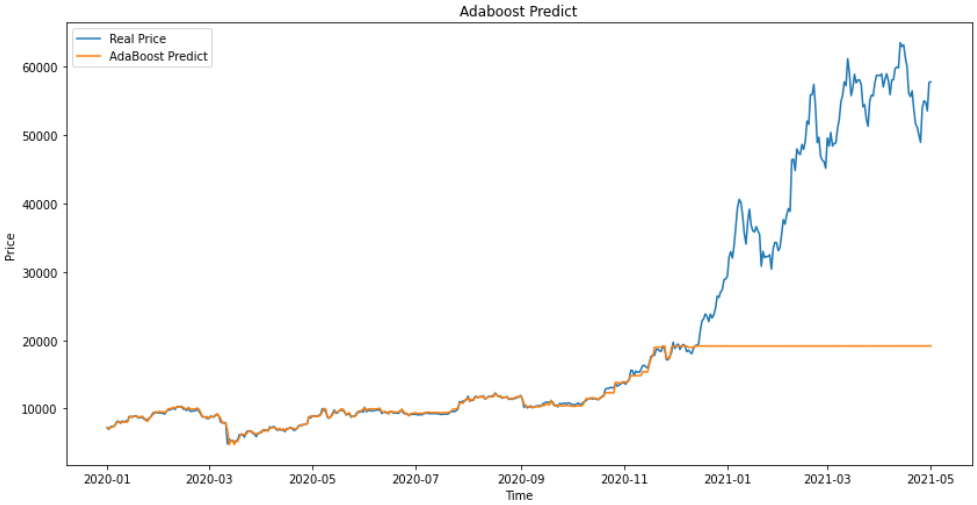
دقت: ۱۵۴۲۰/۶۵ خطا: ۶۹/۴۰

max_depth: ۱۰، n_estimators: ۱۰۰، learning_rate: ۰/۵



دقت: ۱۵۴۱۲/۱۸ خطا: ۶۸/۹۹

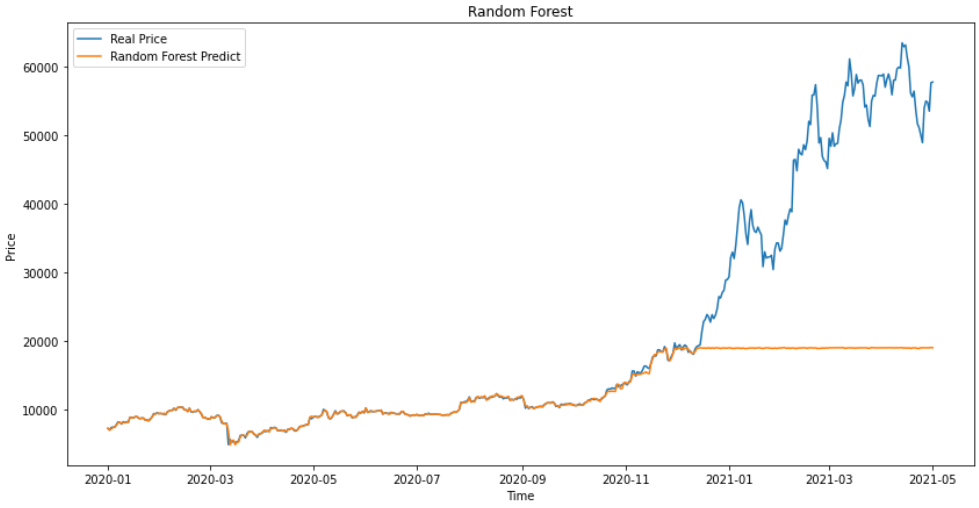
max_depth: ۵، n_estimators: ۱۰۰، learning_rate: ۰/۲



دقت: ۱۵۳۷۴/۹۲ خطا: ۶۸/۱۷

۲۰)

اجرای Random Forest بر روی دیتای بیتکوین



دقت: ۱۵۴۶۷/۰۰ خطا: ۷۰/۸۴