ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Прикладная математика и информатика»

УДК 519.254

Отчет об исследовательском проекте

на тему _____Моделирование временных рядов_____

Выполнил:
студент группы БПМИ188 _____ _____Рябинин А.О_____
Подпись И.О. Фамилия

_____
Дата

Согласовано:
руководитель проекта _____Лукьянченко Петр Павлович_____
Имя, Отчество, Фамилия

_____
Должность / Место работы

Дата _____2020 _____ _____
Оценка (по 10-тибалльной шкале) Подпись

Москва 2020

СОДЕРЖАНИЕ

Введение

В данной работе применяются методы анализа временных рядов для их моделирования и прогнозирования, сравнения результатов, выявления статистически хороших моделей. Моделирование временных рядов широко используется многих сферах анализа и прогнозирования, в этом проекте я буду разбираться с временными рядами стоимостей акций различных фирм, так как это наиболее важная сфера для выявления закономерностей.

На сегодняшний день исследования на данную тему актуальны из-за большого спроса на моделирование временных рядов. Можно прогнозировать многие зависимости от популяций кроликов в различные месяцы года до зависимостей между стоимостью нефти на рынке и ВВП страны.

Объектом исследования являются графики стоимостей акций таких компаний, как Tesla и Yandex, за последний 3 года, с целью прогнозирования будущих значений и оценки различных моделей прогнозирования.

Задачи стоят следующие:

- Заполучить данные за последние 3 года о ранее перечисленных компаниях
- Построить различные модели прогнозирования (SARIMA, Holt, Exponential Smoothing и т.д.)
- Сравнить модели с помощью подсчета среднеквадратичной ошибки, средней абсолютной ошибки, а также средней относительной ошибки

Для исследования и построения моделей использован Python с применением следующих библиотек: pandas, numpy, matplotlib, sklearn, statsmodels.

Основные термины и определения

1. Временной ряд (Time Series) - последовательность измерений, значения которой наблюдаются через равные промежутки времени.
2. Тренд - общая систематическая линейная или нелинейная компонента временного ряда, которая может изменяться о времени.
3. Сезонность - периодически повторяющаяся компонента временного ряда.
4. Стационарный ряд - ряд, обладющий свойством не менять свои характеристики со временем, то есть отсутствуют тренд и сезонность.
5. Модель Скользящего среднего (Moving Average) - модель, значения которой в каждой точке являются некоторым средний значением исходного ряда за предыдущий период
6. Модель одиночного Экспоненциального Сглаживания (Simple Exponentioal Smoothing) - модель, значения которой в каждой точке вычисляются как сумма значений исходного ряда с различными весами. Причем веса уменьшаются экспоненциально, чем ближе к началу ряда, тем меньше вес.
7. Модель SARIMA - модель, объединяющая в себе модель авторегрессии и скользящего среднего.
8. Двухпараметрическая модель Хольта (Holt model) - модель, в которой помимо сглаженных значений считается параметр тренда, таким образом, модель может учитывать тренд.
9. Мульипликативная модель экспоненциального сглаживания Хольта-Уинтерса (Holt-Winters' model) - трехпараметрическая модель, учитывающая и тренд, и сезонность
10. Среднеквадратичная ошибка (MSE) - среднее квадратов ошибки.

11. Средняя абсолютная ошибка (MAE) - среднее абсолютных ошибок.

12. Средняя относительная ошибка (MAPE) - среднее относительных ошибок.

Теоретическая часть

В данной части более подробно разберемся с каждой моделью, а также с методами их оценки, подсчета ошибок.

Объявим элементы временного ряда как $y_1, y_2, \dots y_n$

Прогнозируемые элементы в свою очередь как $S_1, S_2, \dots, S_n, \dots$

1. *Модель скользящего среднего (Moving Average, MA)*

Довольно примитивный, но порой полезный метод прогнозирования. Как уже говорилось ранее, прогнозом является среднее за последние несколько значений исходного временного ряда.

$$S_t = \frac{1}{p}(y_t + y_{t-1} + \dots + y_{t-p+1})$$

Данная модель может быть довольно полезной, если выбрать правильное значение ***p*** для ряда.

2. *Модель Одиночного Экспоненциального сглаживания*

Данная модель отлично представима в виде рекуррентной формулы для прогнозируемых значений.

$$S_t = \alpha * y_{t-1} + (1 - \alpha) * S_{t-1}$$

Параметр **α** выбирается от 0 до 1 с наименьшей среднеквадратичной ошибкой.

Прогноз же в свою очередь будет вычисляться по следующей формуле:

$$F_{t+i} = \alpha * y_t + (1 - \alpha) * S_{t+i-1}$$

Данная модель реализуется в библиотеке **statsmodels** с названием SimpleExpSmothing

3. *Модель SARIMA*

Модель авторегрессии и скользящего среднего. Общая модель, предложенная Боксом и Дженкинсом (1976) включает как параметры авторегрессии, так и параметры скользящего среднего. Именно, имеется три типа параметров модели: параметры авторегрессии (p), порядок разности (d), параметры скользящего среднего (q). В обозначениях Бокса и Дженкинса модель записывается как ARIMA (p, d, q). Например, модель (0, 1, 2) содержит 0 (нуль) параметров авторегрессии (p) и 2 параметра, скользящего среднего (q), которые вычисляются для ряда после взятия разности с лагом 1.

Порядок разности нужен для того, чтобы прогнозировать стационарный ряд. ARIMA является расширенной моделью для модели ARMA, которая отлично прогнозирует в свою очередь стационарные ряды.

Взятие разности позволяет избавиться от тренда временного ряда, для избавления от сезонности требуется еще расширить модель до SARIMA (p, d, q) (P, D, Q, S).

Для модели SARIMA требуется подобрать сезонные параметры.

На практике параметры p(P) и q(Q) редко принимают значение больше 2.

Для нахождения наиболее подходящих параметров для модели SARIMA следует воспользоваться анализом автокорреляционной и частной автокорреляционной функций. В данном же проекте для упрощения будет небольшой перебор по параметрам для вычисления самой подходящей модели с наименьшей среднеквадратичной ошибкой.

4. *Двухпараметрическая модель Хольта*

Данная модель, как и модель одиночного экспоненциального сглаживания выражается рекуррентной формулой для прогнозируемых значений и для параметра тренда b_t

$$S_t = \alpha * y_{t-1} + (1 - \alpha) * (S_{t-1} + b_{t-1})$$
$$b_t = \beta * (S_t - S_{t-1}) + (1 - \beta) * b_{t-1}$$

Параметры **α** и **β** также выбираются от 0 до 1 поиском лучшей модели с наименьшей среднеквадратичной ошибкой.

Прогноз будет вычисляться по следующей формуле

$$F_{t+m} = S_t + m * b_t$$

Данная модель реализуется в библиотеке statsmodels с названием Holt.

5. *Мульипликативная модель экспоненциального сглаживания Хольта-Уинтерса*

Данная модель также выражается рекуррентной формулой с параметрами **α**, **β**, **μ,** где параметр $\mathbf{I_t}$ отвечает за сезонность. ***L*** отвечает за период сезонности.

$$S_t = \alpha * \frac{y_t}{I_{t-L}} + (1 - \alpha) * (S_{t-1} + b_{t-1})$$
$$b_t = \beta * (S_t - S_{t-1}) + (1 - \beta) * b_{t-1}$$
$$I_t = \mu * \frac{y_t}{S_t} + (1 - \mu) * I_{t-L}$$

Прогноз будет вычисляться по следующей формуле

$$F_{t+m} = (S_t + m * b_t) * I_{t-L+m}$$

Данная модель реализуется в библиотеке statsmodels с названием ExponentialSmoothing.

6. *Подсчет Ошибок*

В данной части разберемся с оценкой моделей и подсчетом различных ошибок для их сравнения

- Среднеквадратичная ошибка (MSE)

$$MSE = \frac{1}{n}\sum(S_t - y_t)^2$$

- Средняя абсолютная ошибка (MAE)

$$MAE = \frac{1}{n}\sum|y_t - S_t|$$

- Средняя относительная ошибка (MAPE)

$$MAPE = \frac{1}{n}\sum\frac{|y_t - S_t|}{y_t}$$

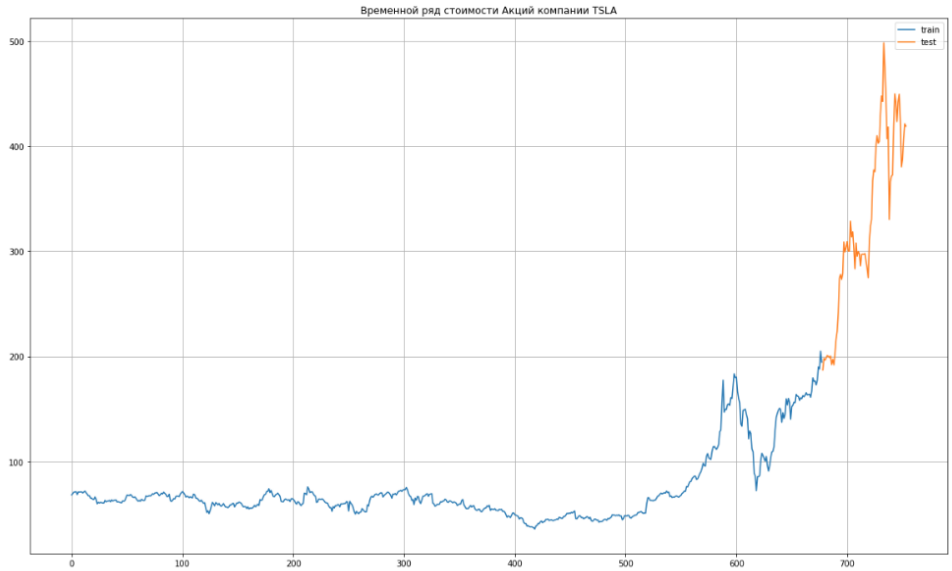
Описание вычислительного эксперимента

Эксперимент проводился с помощью jupyter notebook для наглядности результата. Загружаем данные с помощью “*.csv”. Последовательно строим прогноз для каждой модели, вычисляем ошибки, строим график прогноза. Далее выбираем лучшую модель.

1. **Tesla**

Данные загружены с сайта finance.yahoo.com.

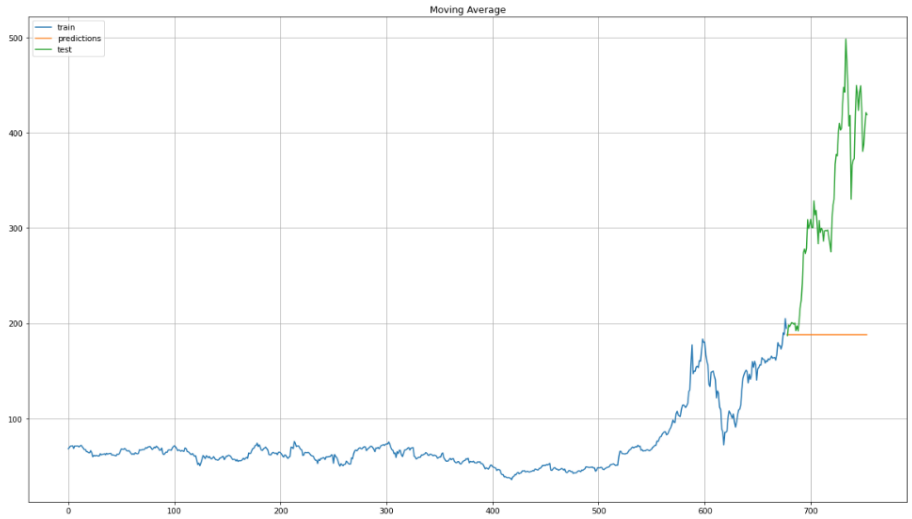
На графике представлены тренировочная и тестовая выборки



На графике четко виден тренд из-за которого многие модели сработают плохо.

1.1. Moving Average

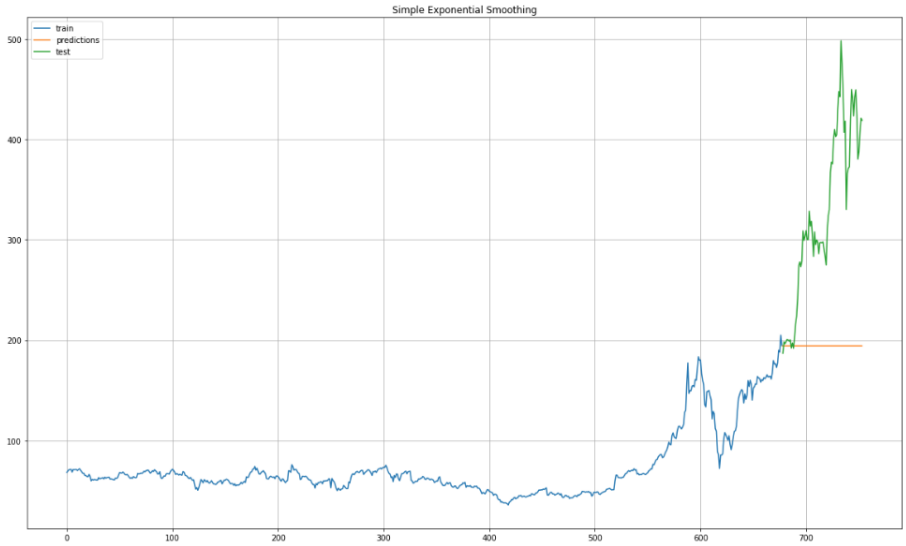
Данная модель показывает не лучший результат, так как она совсем не учитывает тренд. Следует сначала от него избавиться и анализировать стационарный ряд (от сезонности стоит тоже избавиться).




```
MSE =  26240.627686806583
MAE =  138.6468498421052
MAPE =  0.3807880975336937
```

1.2. Simple Exponential Smoothing

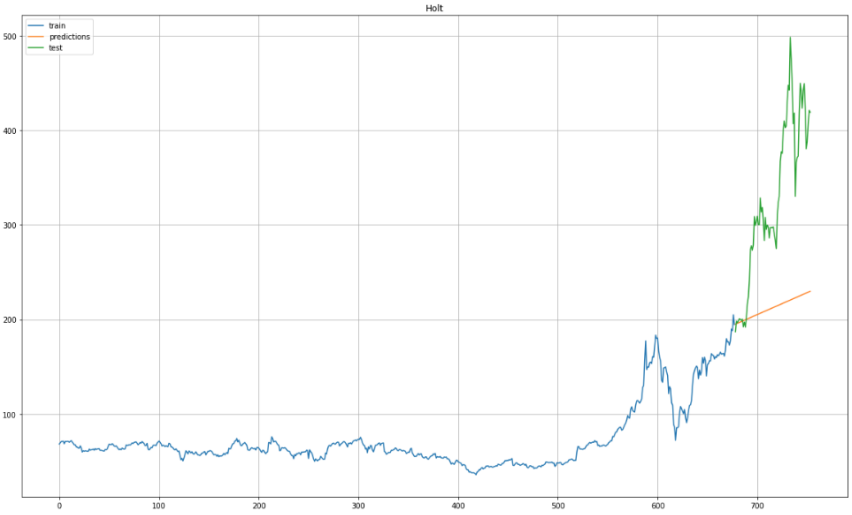
Данная модель также не может справиться с трендом и сезонностью, как уже говорилось ранее, тем самым получается большая ошибка, как в модели скользящего среднего.




```
MSE =  24449.785171105683
MAE =  132.33571525142293
MAPE =  0.3606011163608339
```

1.3. Holt

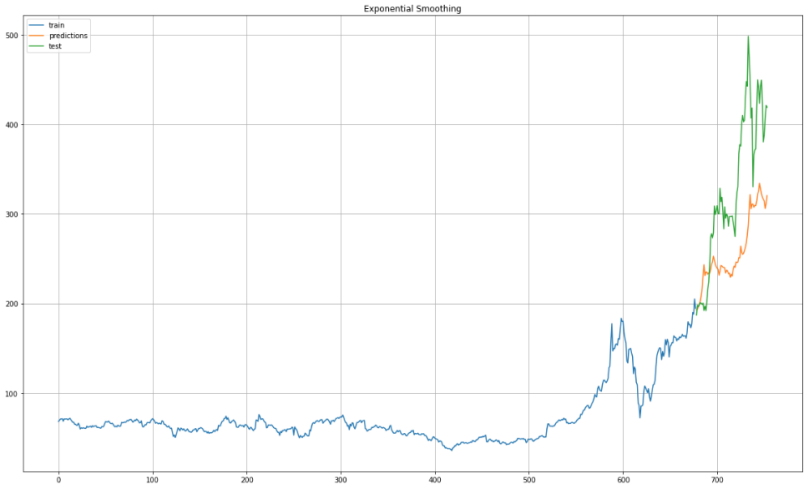
Данная модель уже может прогнозировать тренд, но успех не очень большой. Сезонность данная модель не учитывает. Получаем меньшую ошибку, но результат неудовлетворительный.




```
MSE =  18613.180268008484
MAE =  114.78212624672443
MAPE =  0.3118946343164158
```

1.4. Exponantial Smoothing

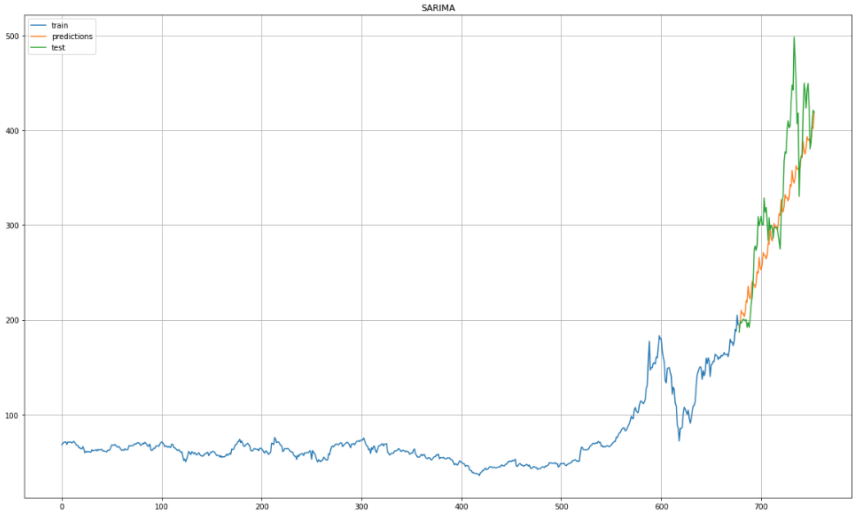
Модель Хольта-Уинтерса уже с большим успехом справляется с трендом и сезонностью. Данная модель хорошо подходит для прогнозирования временных рядов.




```
MSE =  7785.306316842169
MAE =  74.34420102696994
MAPE =  0.2080890911453014
```

1.5. SARIMA

В этом эксперименте данная модель показала наилучший результат. Очень маленькая средняя относительная ошибка говорит о том, что модель можно применять на практике гораздо чаще, чем остальные.



```
MSE =  2142.631204754725
MAE =  34.30141834217948
MAPE =  0.09807301487831249
```

1.6. Сравнение результатов, выбор лучшей модели

Модели скользящего среднего и одиночного экспоненциального сглаживания спрогнозировали плохо. Очевидно, что это происходит из-за того, что они не учитывают тренд. Обе модели дали среднюю относительную ошибку около 36-38%. Модель Хольта справилась чуть лучше, благодаря возможности уловить тренд, но средняя относительная ошибка стала несильно меньше - 31%. Модели, учитывающие сезонность, показали себя лучше всех. Модель Хольта-Уинтерса добилась средней относительной ошибки в 20%, модель SARIMA – 9%.

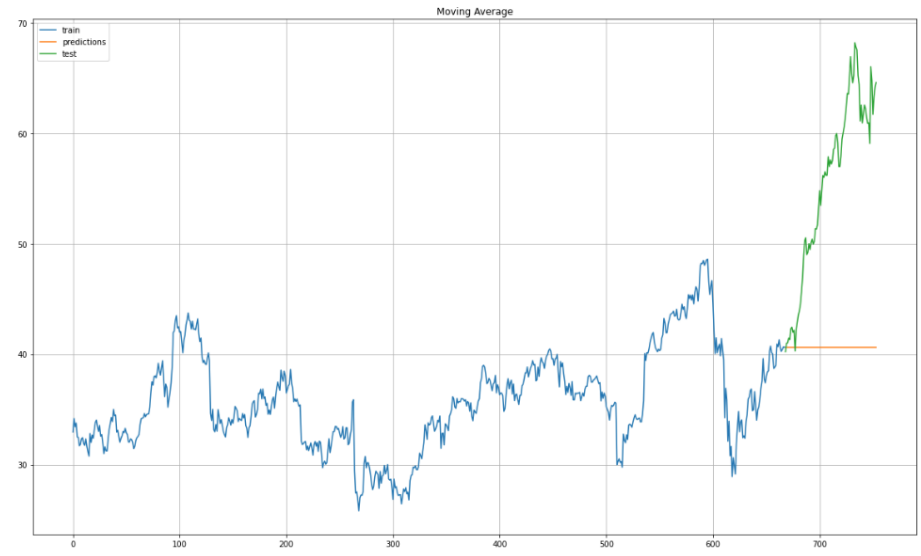
В таблице ниже можно увидеть результаты всех ошибок для каждой модели.

	MA	SimpleExpSm	SARIMA	Holt	ExpSm
MSE	26240.6	24449.8	2142.63	18613.2	7785.31
MAE	138.647	132.336	34.3014	114.782	74.3442
MAPE	0.380788	0.360601	0.098073	0.311895	0.208089

2. Yandex

2.1. Moving Average

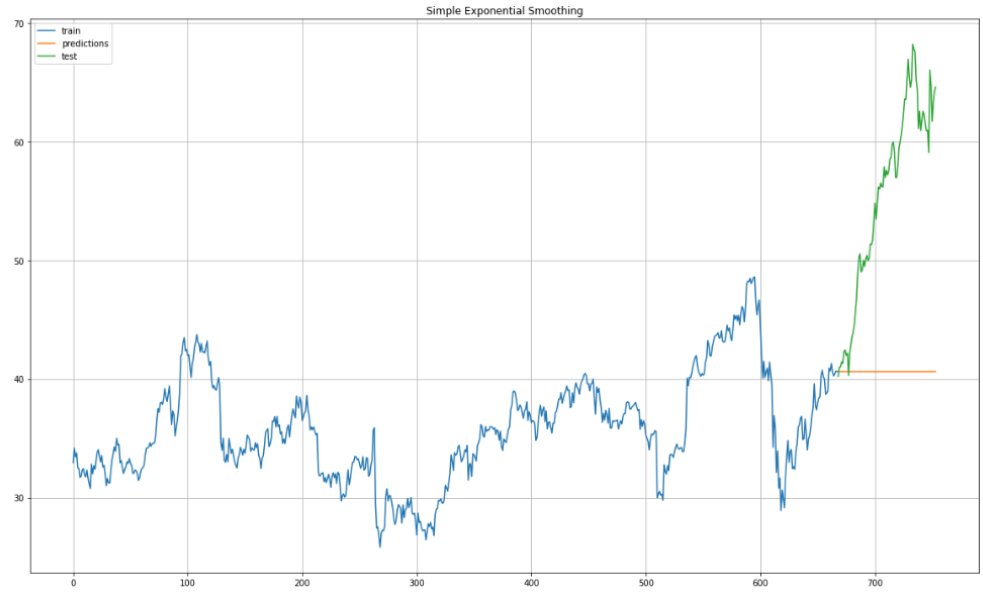
Как и в случае с компанией TESLA, данная модель также не в силах справиться с трендом и сезонностью, из-за чего получается большая ошибка.




```
MSE =  285.02735313407504
MAE =  14.833606430232564
MAPE =  0.25017369393191174
```

2.2. Simple Exponential Smoothing

В данной модели получается аналогичная ситуация, когда прогноз далек от реальности из-за влияния тренда и сезонности.

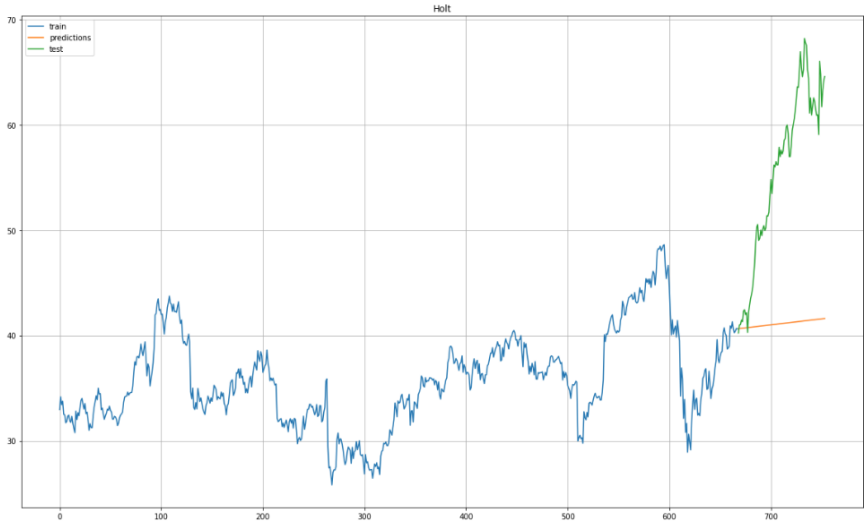



```
MSE =  285.4465609876577
MAE =  14.847089819428952
MAPE =  0.2504183293878882
```

2.3. Holt

В очередной раз модель Хольта даёт неудачный прогноз.

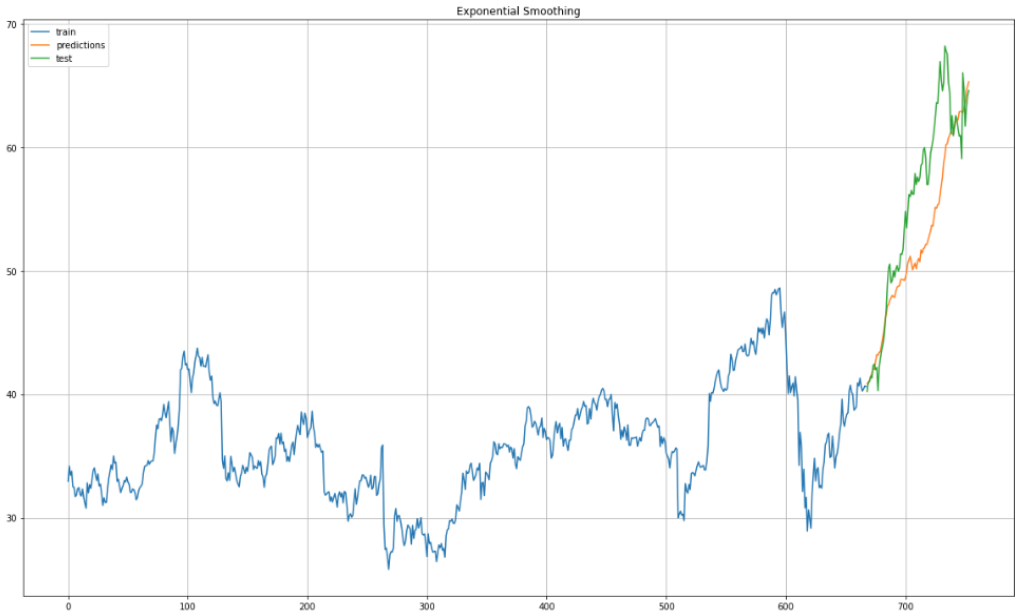
Модель не в состоянии уловить тренд, а также вовсе не учитывает сезонность. Можно сделать вывод, что реализация данной модели в библиотеке statsmodels требует доработок.




```
MSE =  266.76621586653636
MAE =  14.353661751977087
MAPE =  0.24209965282245757
```

2.4. Exponantial Smoothing

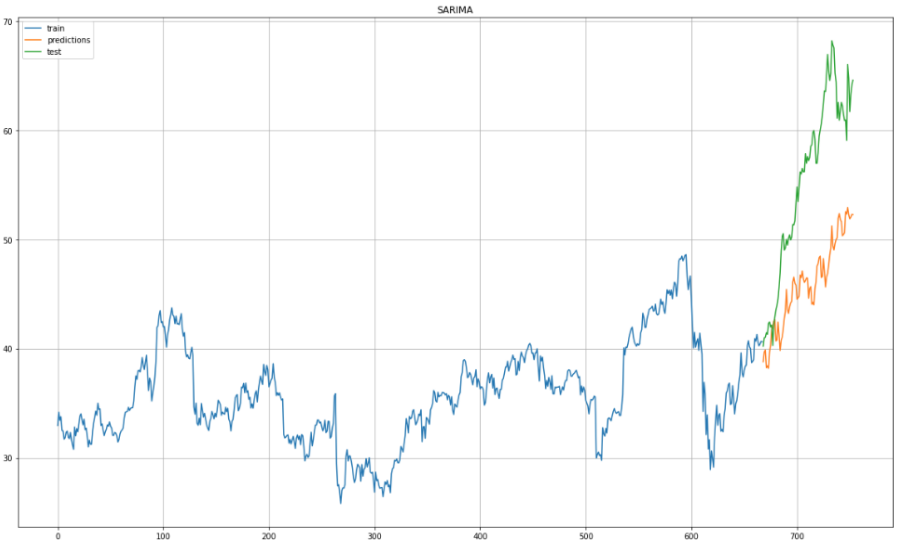
В свою очередь модель тройного экспоненциального сглаживания показала более лучший результат, чем с акциями TESLA. Модель хорошо уловила тренд и сезонность, сделав хороший прогноз.




```
MSE =  22.054984085655608
MAE =  3.648679100890017
MAPE =  0.06215601404072724
```

2.5. SARIMA

Модель SARIMA же показала результат хуже, чем прошлый раз, ошибка сильно превышает среднюю относительную ошибку модели Хольта-Уинтерса. Хороши обе модели, но стоит их правильно выбирать.




```
MSE =  115.5377282325806
MAE =  9.531862383763551
MAPE =  0.16262198491586252
```

2.6. Сравнение результатов, выбор лучшей модели

Первые две модели в очередной раз показали не очень хороший результат со средней относительной ошибкой в 25%. Двухпараметрическая модель Хольта уловила тренд, но значение средней относительной ошибки несильно меньше, чем у моделей, которые тренд не в состоянии отследить. Средняя относительная ошибка третей модели - 24%. Модель Хольта-Уинтерса показала наилучший результат со средней относительной ошибкой в 6%, в то время как модель SARIMA дала ошибку в 16%. В таблице ниже приведены все ошибки для каждой модели.

	MA	SimpleExpSm	SARIMA	Holt	ExpSm
MSE	285.027	285.447	115.538	266.766	22.055
MAE	14.8336	14.8471	9.53186	14.3537	3.64868
MAPE	0.250174	0.250418	0.162622	0.2421	0.062156

Заключение

Рассмотрев пять предложенных моделей для прогнозирования временного ряда стоимостей акций компаний TESLA и YANDEX, можно сделать несколько выводов.

- Модели скользящего среднего и одиночного экспоненциального сглаживания дают плохой результат на нестационарных рядах. Поэтому следует сначала сделать ряд стационарным, а далее уже применять данные модели для прогнозирования. Преимущество этих моделей заключается в том, что они имеют несложную реализацию.
- Модель Хольта имеет много недостатков, по сложности же она несильно отличается от модели Хольта-Уинтерса. Средняя относительная ошибка прогноза этой модели не была меньше 20%, что не является желаемым результатом.
- Модели Тройного Экспоненциального Сглаживания и SARIMA дают наилучший результат при прогнозировании временных рядов. Эти модели следует использовать в приоритете над остальными моделями. Основной недостаток моделей заключается в сложности реализации, подборе параметров и времени работы.

Анализ временных рядов позволяет выделить компоненты ряда - тренд и сезонность. Это позволяет давать точные прогнозы.

Список использованных источников

[1] Engineering Statistics Handbook (2012), Introduction to Time Series Analysis. Available: https://itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm

[2] Электронный учебник по статистике. Анализ временных рядов Available: http://statsoft.ru/home/textbook/modules/sttimser.html#exponential

[3] Statsmodels documentation. Available: https://www.statsmodels.org/dev/index.html

[4] Сайт http://www.machinelearning.ru/, статья “Временной Ряд”

[5] В.Н. Афанасьев, М.М. Юзбашев “Анализ временных рядов и прогнозирование”

[6] Hyndman R. J., Athanasopoulos G. “Forecasting principles and practice"