Итоговая аттестационная работа

Тема для исследования:

«Анализ и прогнозирование временных рядов с дальнейшей оценкой точности выбранных моделей на примере данных о продажах новых машин в Норвегии»

1. Знакомство с данными

Для работы были взяты данные из открытого источника
https://www.kaggle.com/datasets/dmi3kno/newcarsalesnorway

Данные представляют собой месячное количество проданных новых автомобилей в Норвегии в разрезе автомобильных марок за период с 2007 по 2017г.

Рис. 1 Результат вывода данных средствами pandas

	Year	Month	Make	Quantity	Pct
0	2007	1	Toyota	2884	22.7
1	2007	1	Volkswagen	2521	19.9
2	2007	1	Peugeot	1029	8.1
3	2007	1	Ford	870	6.9
4	2007	1	Volvo	693	5.5
...	...	...	...	...	...
4372	2017	1	Nilsson	3	0.0
4373	2017	1	Maserati	2	0.0
4374	2017	1	Ferrari	1	0.0
4375	2017	1	Smart	1	0.0
4376	2017	1	Ssangyong	1	0.0

4377 rows × 5 columns

Целью работы являлись проведение анализа временного ряда и дальнейшее его прогнозирование с помощью моделей SARIMA, Prophet, ETS Model (Exponential Smoothing).

Для этих целей данные были сгруппированы в соответствии с марками автомобилей. Результаты были сравнены и выбран производитель с наибольшим числом продаж за рассматриваемый период. Сформирован дата-фрейм для дальнейшего анализа временного ряда и построены соответствующие графики.

Рис. 2 Секторная диаграмма по числу продаж автомобилей с 2007 по 2017г

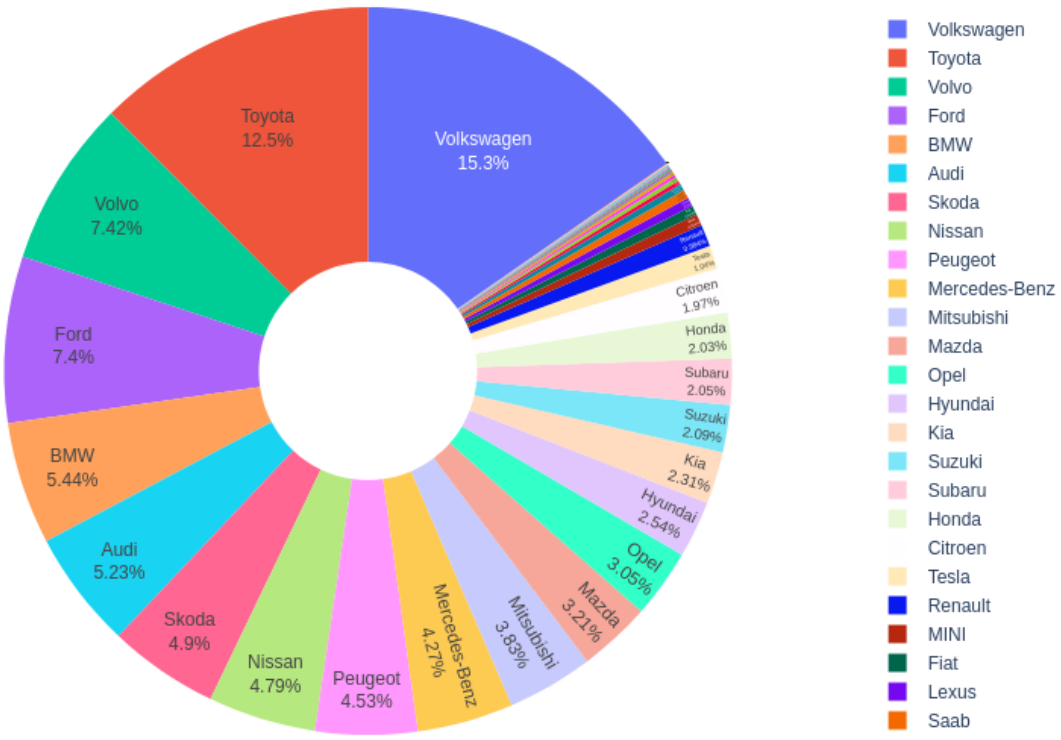
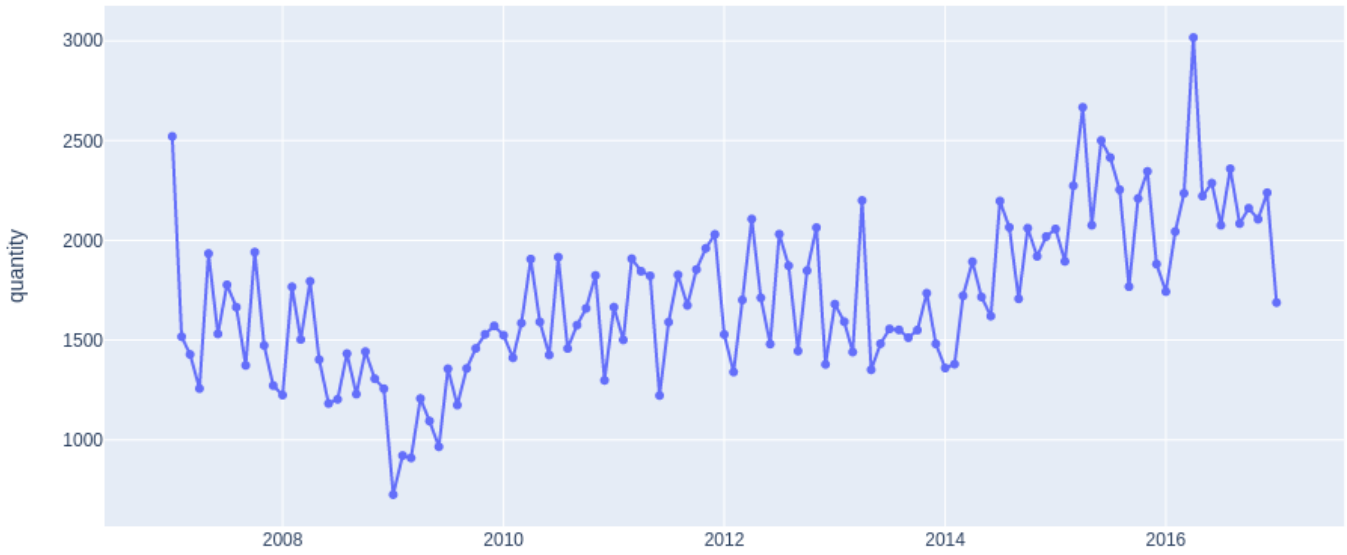


Рис. 3 Общий график временного ряда



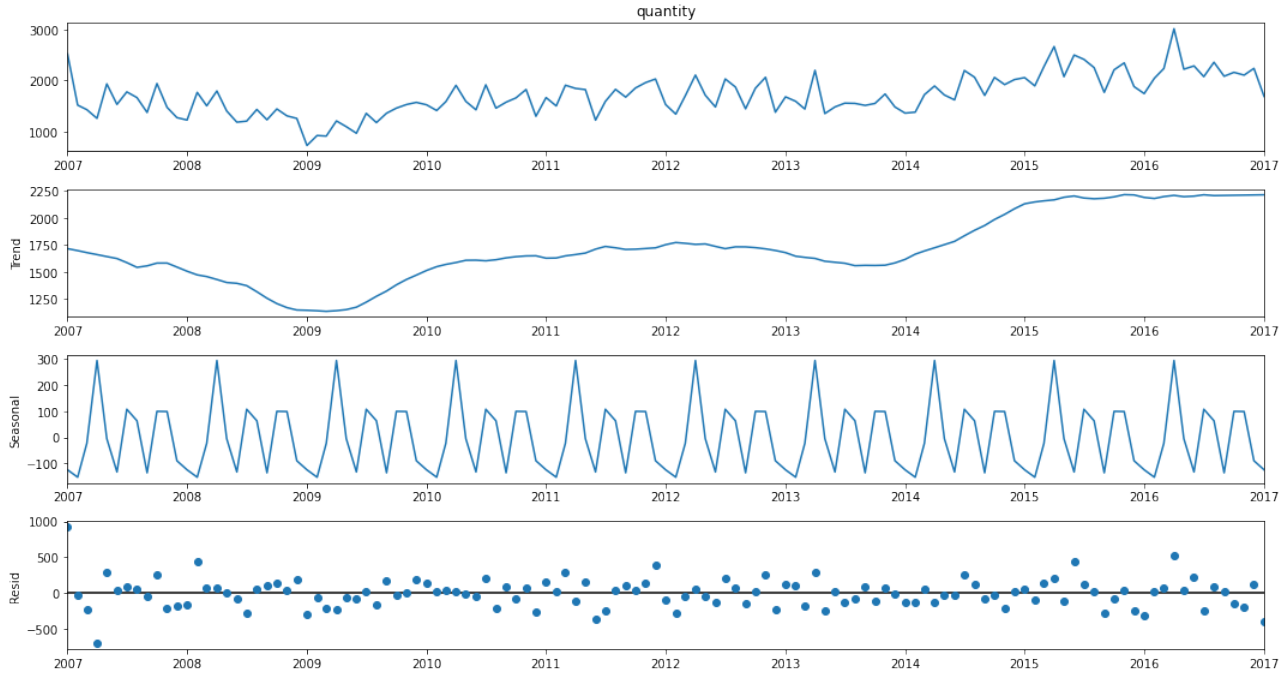
2. Обработка данных

С первого взгляда на графике трудно определить чёткую сезонность и ярко - выраженный тренд, а также стационарность ряда.

Для определения стационарности была применена комбинация тестов Дики — Фуллера и Квятковского-Филлипса-Шмидта-Шина. В результате были получены значения, указывающие на необходимость извлечения трендовой составляющей из временного ряда для удовлетворений требованиям о стационарности.

Далее была выполнена сезонная декомпозиция временного ряда и построены графики функций автокорреляции и частичной автокорреляции.

Рис. 4 Составляющие сезонной декомпозиции временного ряда



Из графиков видно, что тренд распределён нелинейно. Имеются частичные выравнивания функции в средней и конечной частях графика. График сезонной составляющей имеет определённую цикличность.

Были построены графики автокорреляции исходных данных, а также данных с последовательным извлечением составляющих тренда и сезонности для установления их влияния на стационарность ряда. Остатки также были проверены на стационарность и оказались «белым шумом», что является хорошим показателем и означает, что функция декомпозиции выбрала данные должным образом.

Для уменьшения дисперсии временного ряда было применено преобразование Бокса-Кокса.

Рис. 5 Графики автокорреляции исходных данных ряда

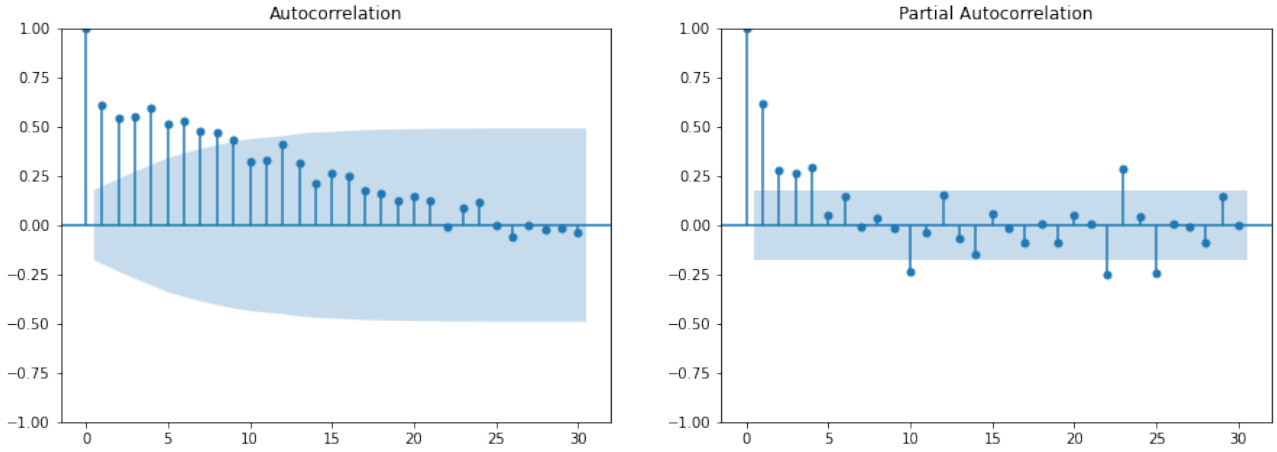


Рис. 6 Графики автокорреляции после извлечения сезонной составляющей

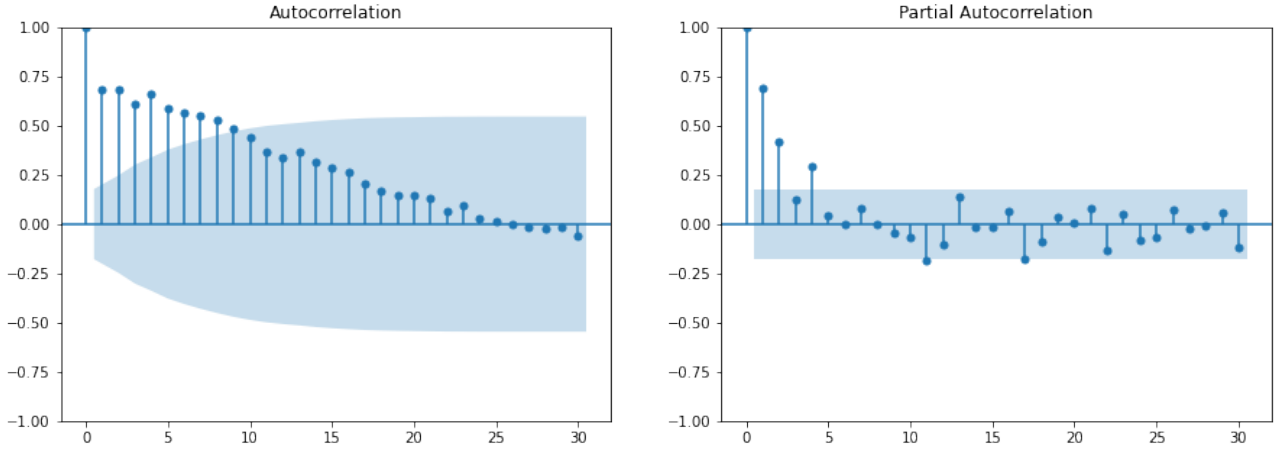
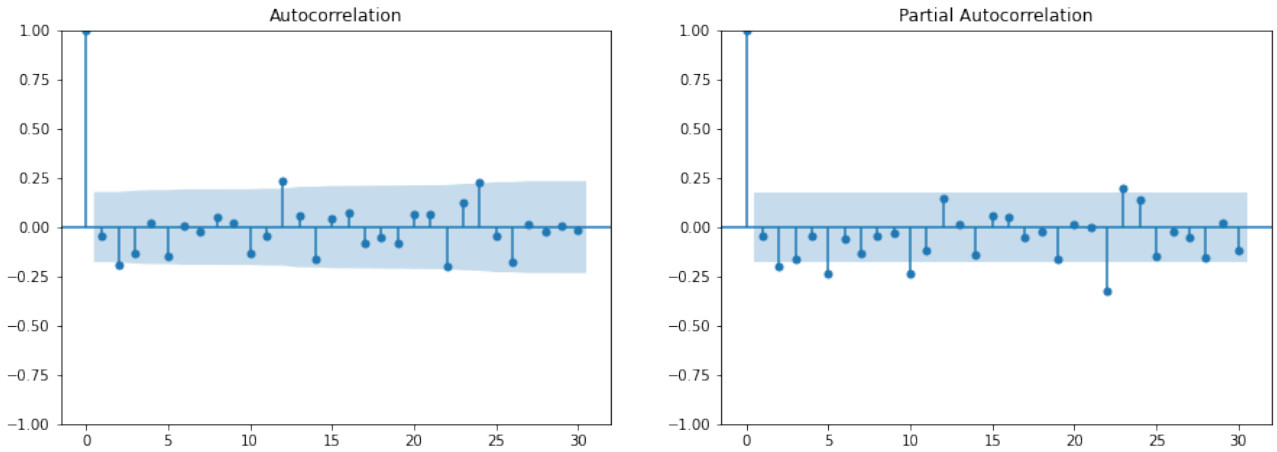


Рис. 7 Графики автокорреляции после извлечения трендовой составляющей



Из графиков выше видно, что сезонность не оказывает особого влияния на стационарность нашего временного ряда, в то время как извлечение тренда приводит ряд к стационарному виду.

Таким образом, было принято решение выделять трендовую составляющую ряда и использовать её в дальнейшем как экзогенный фактор при построении модели SARIMA.

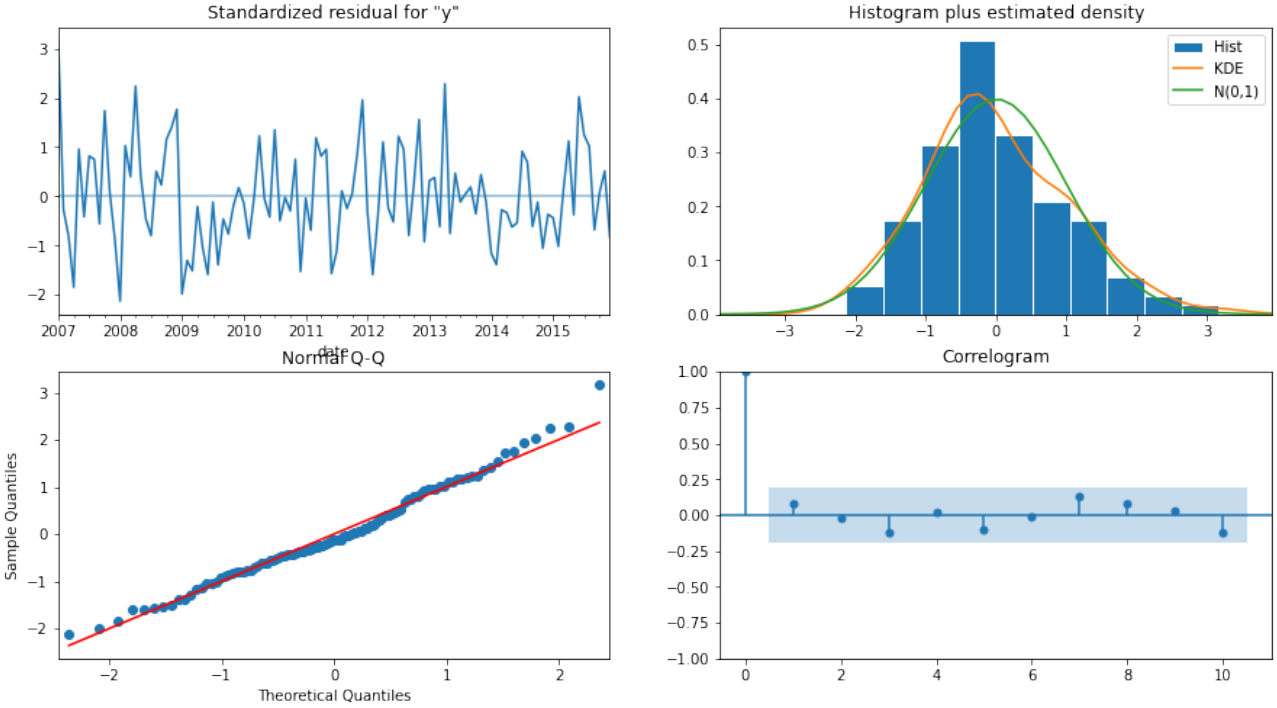
3. Построение моделей и прогнозирование

3.1 SARIMA

Для подбора параметров к данной модели была использована функция auto_arima(). В результате была подобрана модель вида SARIMAX(1, 0, 1)x(1, 0, 1, 12).

После обучения модели на тренировочной и полном дата-фреймах получились следующие результаты.

Рис. 8 Оценочные показатели обучения модели на тренировочном наборе данных



Из графиков выше видно, что диагностические данные имеют распределение, близкое к нормальному. Ярко — выраженной сезонности нет, имеются показатели стационарности. Данные выводы свидетельствуют об удовлетворительных результатах обучения нашей модели.

Рис. 9 Работа модели на тестовом наборе данных

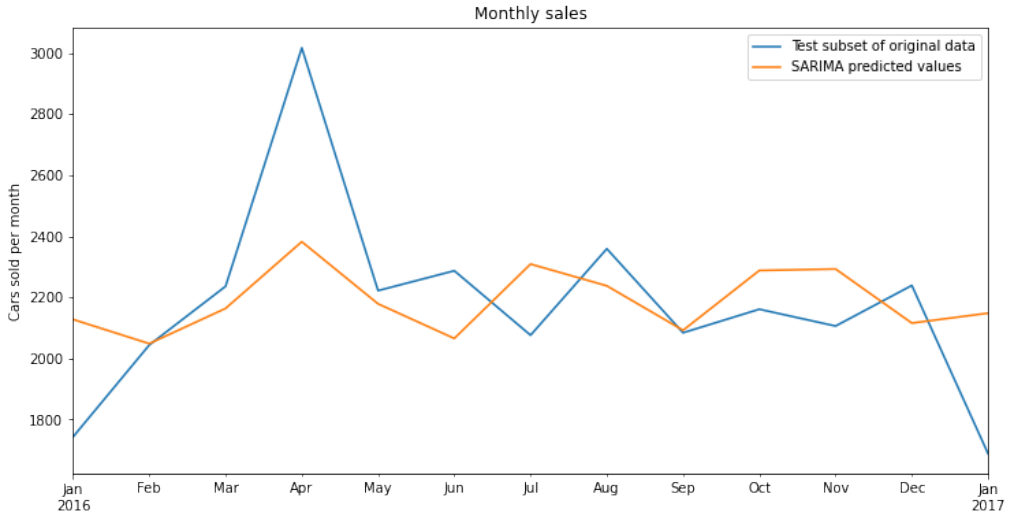
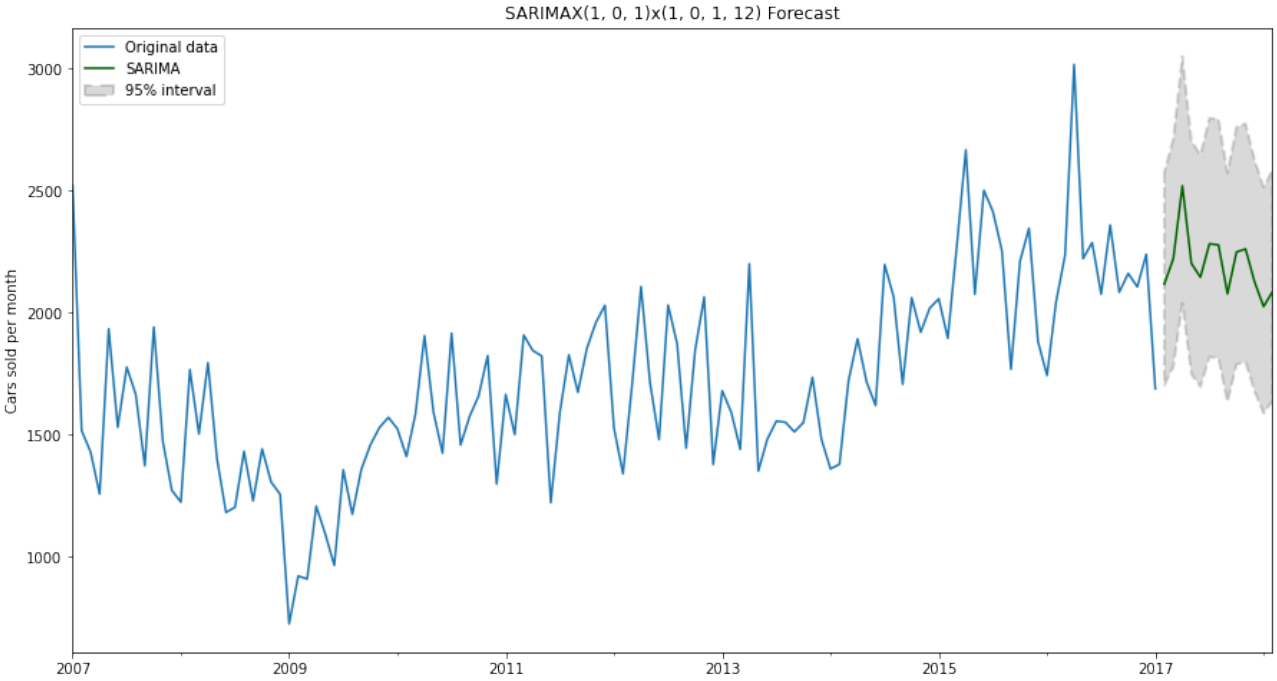


Рис. 10 Прогнозирование на год вперёд после обучения модели на полном наборе данных



Выводы:

Модель показала неплохие результаты на тестовых данных со следующими метриками.

MAPE: 0.094202
ME: 14.40062899356868
MAE: 201.59130077466136
MPE: 0.022560381748378846
RMSE: 270.92512123168655

Таким образом, мы видим, что уровень точности оказался в пределах 90% при прогнозировании данных на год вперёд, что является хорошим показателем. Основные расхождения между фактическими и прогнозируемы данными связаны с наличием пиков спада и подъёма продаж в данных, которые не попали в тренировочный набор данных (начало 2016г.)

3.2 Prophet

В качестве второй модели для прогнозирования была взята модель Prophet. Модель не требовательна к предобработке данных. После приведения данных к необходимому формату, модель также была обучена на тренировочном и полном наборах данных. Результаты были получены близкие к тем, которые мы наблюдали при работе с SARIMA.

Рис. 11 Работа модели на тестовом наборе данных

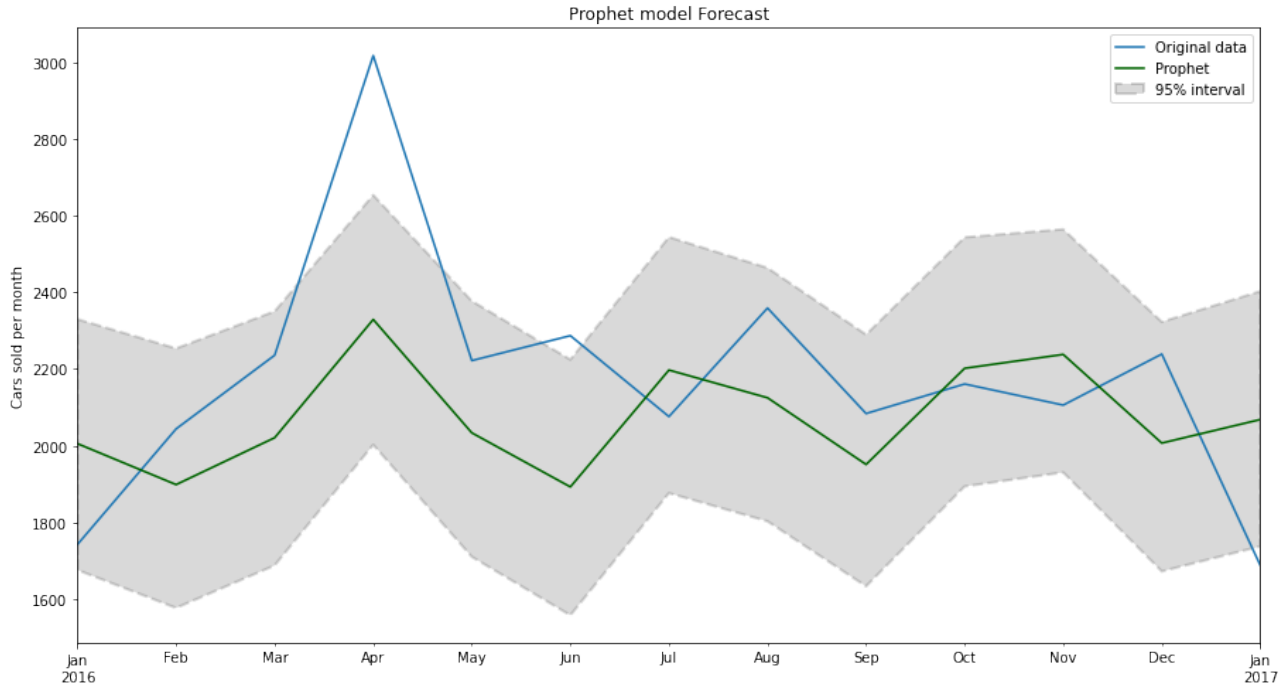
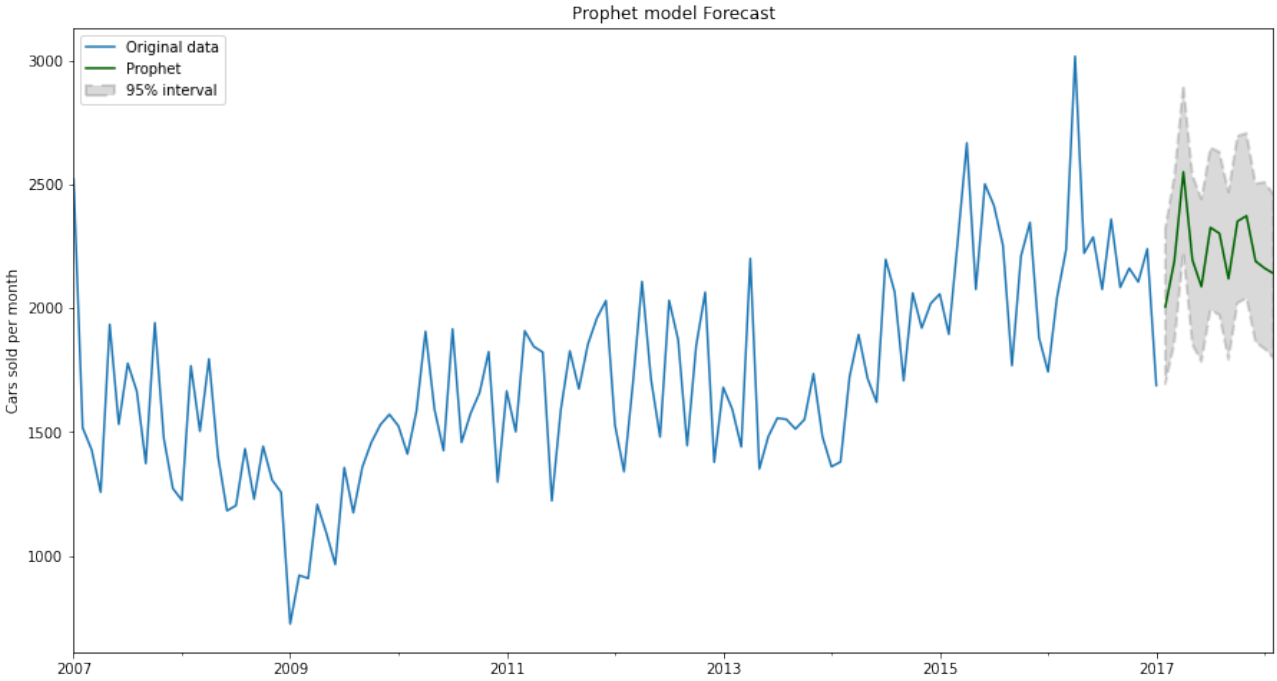


Рис. 12 Прогнозирование на год вперёд после обучения модели на полном наборе данных



Выводы:

Модель показала неплохие результаты на тестовых данных со следующими метриками.

MAPE: 0.110369
ME: -99.3746532375908
MAE: 243.56792025110548
MPE: -0.030973862750652194
RMSE: 291.5323569178151

Уровень точности оказался в пределах 89% при прогнозировании данных на год вперёд, что является хорошим показателем. Результаты оказались близки к данным, полученным в результате прогнозирования, используя модель SARIMA.

3.3 ETS Model

Финальной была выбрана модель экспоненциального сглаживания с настройками тренда, ошибки и сезонности. Шаги с обучением модели были проведены как и в примерах выше.

Рис. 13 Работа модели на тестовом наборе данных

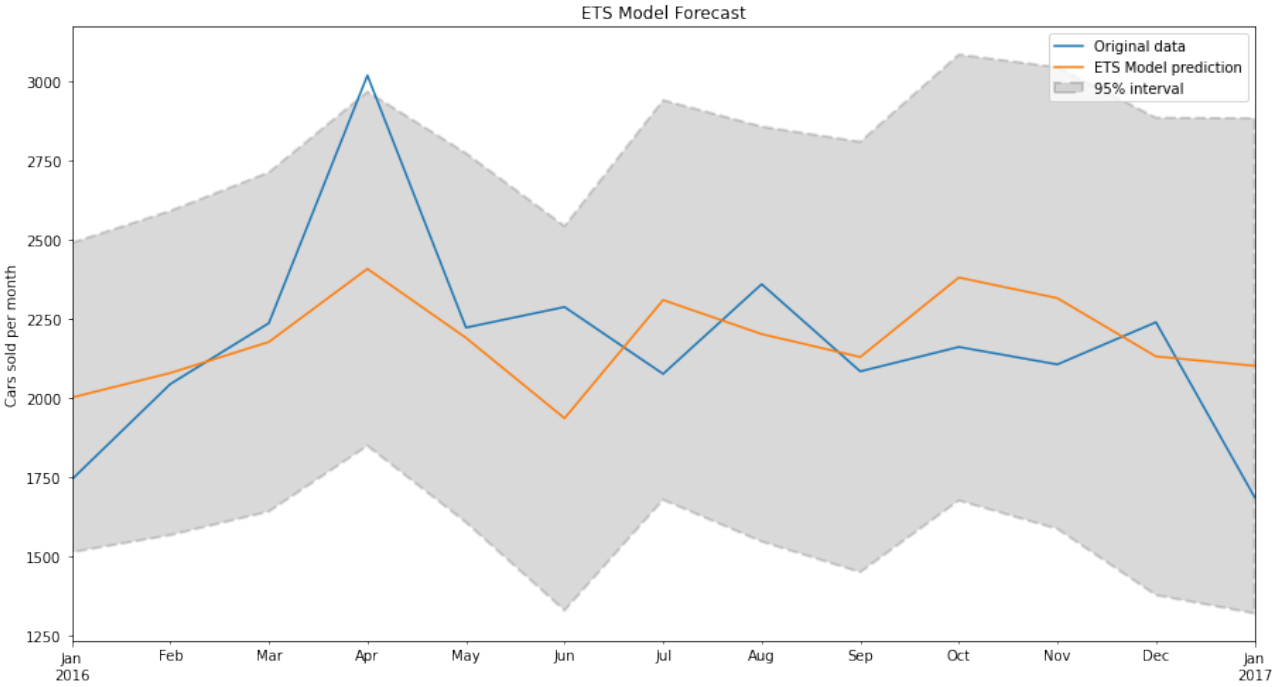
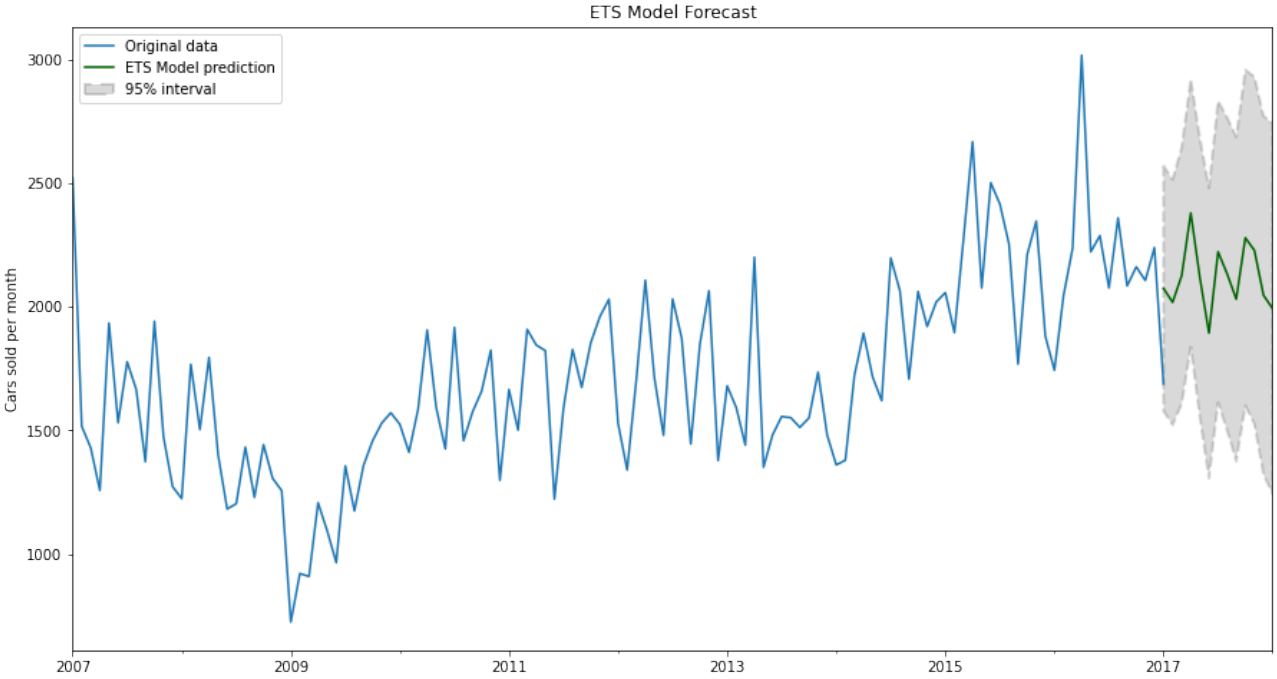


Рис. 14 Прогнозирование на год вперёд после обучения модели на полном наборе данных



Результаты прогнозирования третьей модели не показали ярко - выраженных отличий от результатов, полученных моделями SARIMA и Prophet.

Метрики:

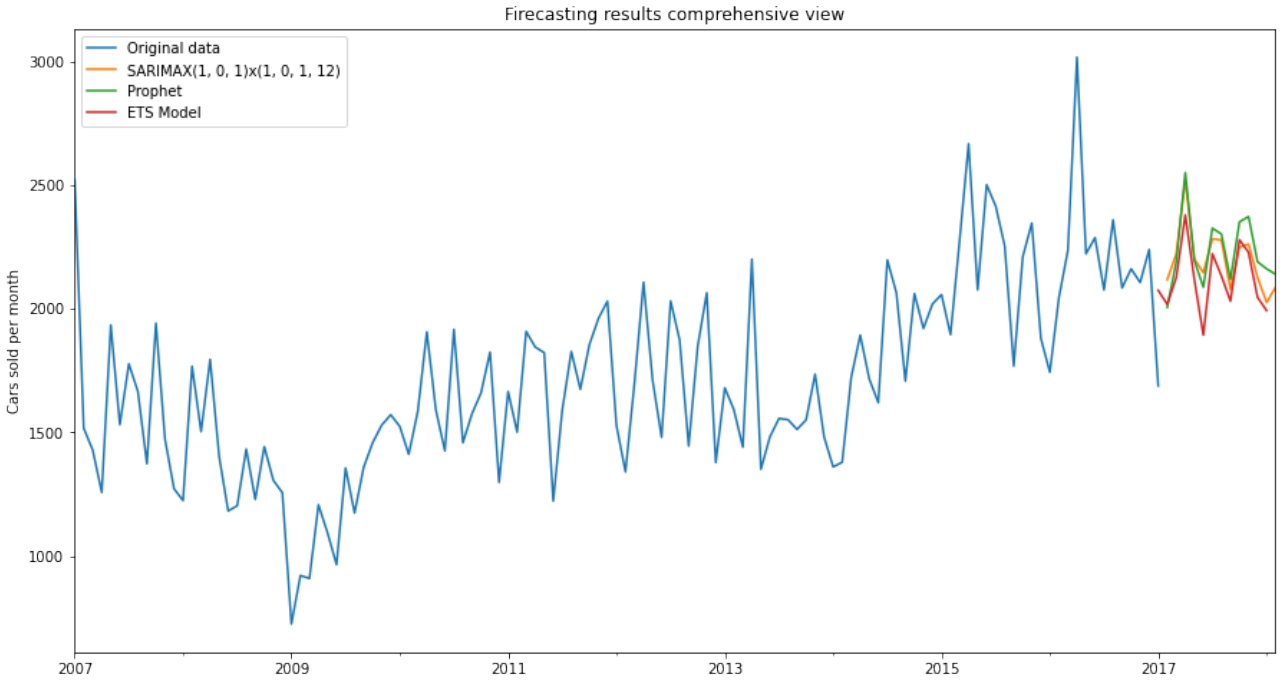
MAPE: 0.096626
ME: 7.372648888730999,
MAE: 210.00649485463015,
MPE: 0.017959551907327256,
RMSE: 266.3963811129286

Уровень точности оказался в пределах 90% при прогнозировании данных на год вперёд, что является хорошим показателем.

4. Общие выводы по работе

Наконец, сравнив метрики всех трёх моделей, был сделан вывод, что все они имеют более или менее одинаковый уровень точности при предсказании одного и того же количества наблюдений.

Рис. 15 Сводный график прогнозов рассматриваемых моделей



По - видимому, значительные значения RMSE и MAE были вызваны всплеском в исходных данных в начале 2016 года, который не следовал ни трендовым, ни сезонным закономерностям. В остальном, подобранные модели неплохо уловили общее поведение временного ряда.

Упомянутый в начале тестовых данных всплеск мог быть вызван сезонными распродажами у дилеров или схожими факторами. Также, в описательной части исходного набора данных

было упомянуто, что спад продаж автомобилей с ДВС в определённое время был вызван ростом продаж электрокаров.

И хотя показатели всех моделей почти одинаковы, модели ETS и SARIMA (Auto_arima) показали немного лучшую производительность. Однако модель ETS требует значительно меньше времени и не требует столько манипуляций с данными, как SARIMA.
Все модели могут быть усовершенствованы для повышения производительности после более тщательного анализа исходных данных.

Таб. 1 Сводная таблица метрик

Модель	MAPE	ME	MAE	MPE	RMSE
SARIMA	0.094202	14.400629	201.591301	0.022560	270.925121
Prophet	0.110369	-99.374653	243.567920	-0.030974	291.532357
ETS Model	0.096626	7.372649	210.006495	0.017960	266.396381