Министерство образования Республики Беларусь

Учреждение образования
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет Компьютерных сетей и систем

Кафедра Информатики

МАШИННОЕ ОБУЧЕНИЕ

ЛАБОРАТОРНАЯ РАБОТА №8
«Рекуррентные нейронные сети для анализа временных рядов»

БГУИР 1-40 81 04

Магистрант:
гр. 858641
Кукареко А.В.

Проверил:
Стержанов М. В.

Минск, 2020

ХОД РАБОТЫ

Данные.

Набор данных для прогнозирования временных рядов, который состоит из среднемесячного числа пятен на солнце, наблюдаемых с января 1749 по август 2017. Данные в виде csv-файла можно скачать на сайте Kaggle -> https://www.kaggle.com/robervalt/sunspots/ .

Задание.

1. Загрузите данные. Изобразите ряд в виде графика. Вычислите основные характеристики временного ряда (сезонность, тренд, автокорреляцию);
2. Для прогнозирования разделите временной ряд на обучающую, валидационную и контрольную выборки.
3. Примените модель ARIMA для прогнозирования значений данного временного ряда.
4. Повторите эксперимент по прогнозированию, реализовав рекуррентную нейронную сеть (с как минимум 2 рекуррентными слоями).
5. Сравните качество прогноза моделей.

Результат выполнения:

1. Загрузите данные. Изобразите ряд в виде графика. Вычислите основные характеристики временного ряда (сезонность, тренд, автокорреляцию).

Ряд в виде графика представлен на рисунке 1.

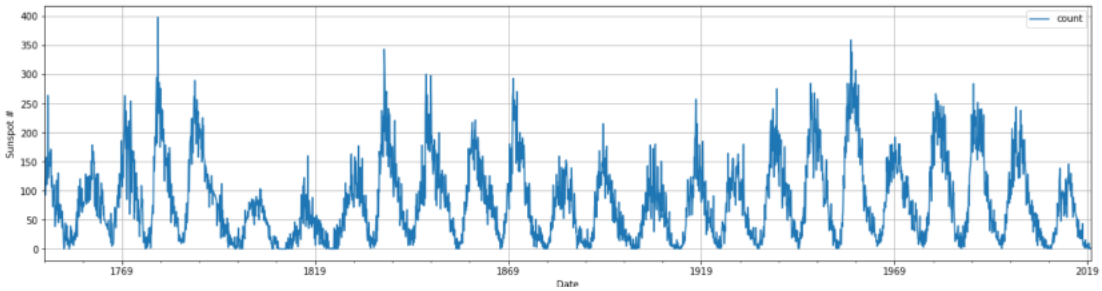

Рисунок 1 – график среднемесячного числа пятен на солнце.

Если масштабировать график, то можно заметить наличие некоторой сезонности. Пример можно увидеть на рисунке 2.

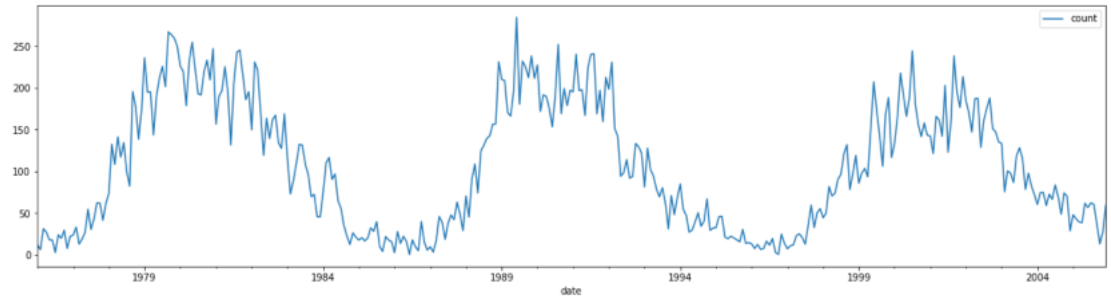

Рисунок 2 – график среднемесячного числа пятен на солнце с 1976 по 2006 г.

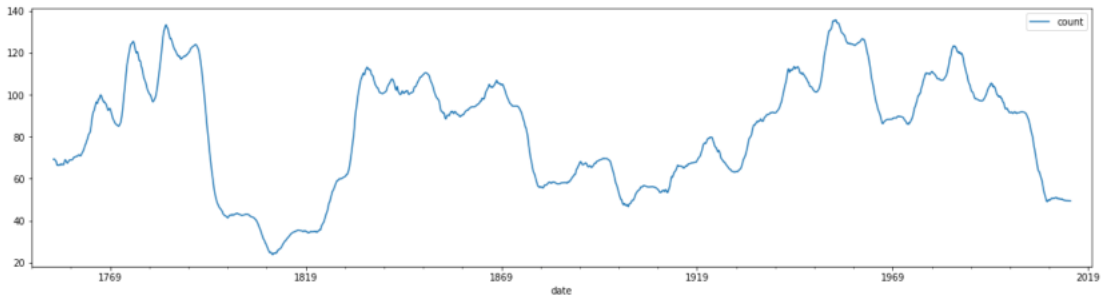

Рисунок 4 – тренда временного ряда.

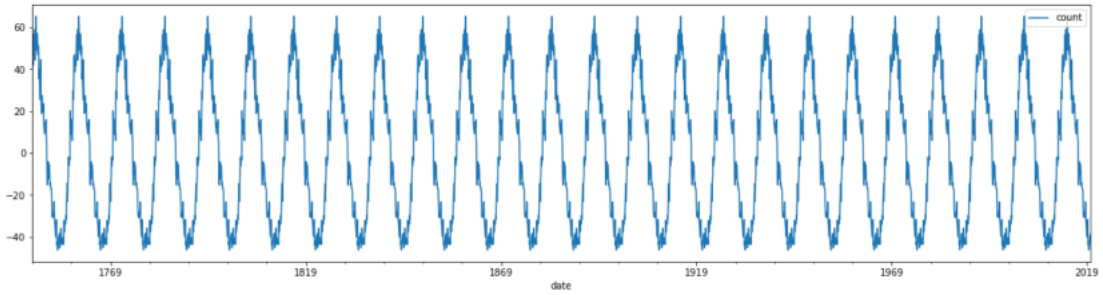

Рисунок 5 – сезонность временного ряда.

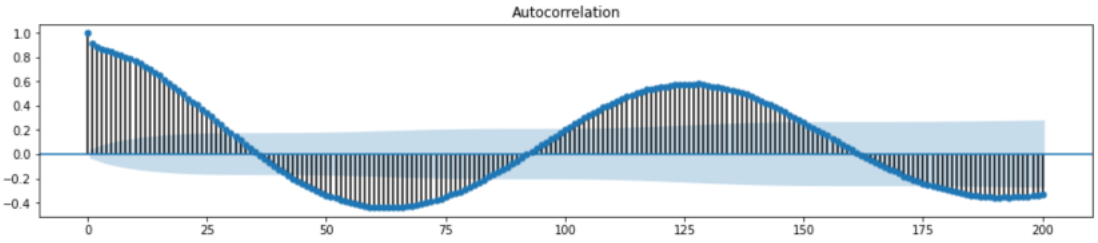

Рисунок 6 – автокорреляция временного ряда.

По графикам, можно увидеть, что временной ряд имеет сезонность и она приблизительно равна 11 годам (132 месяцам).

Так же временной ряд был проверен на наличие стационарности методом «Дики - Фуллера».

Результаты проверки:

- p-value - 1.137033e-18 (очень маленькое).

По результатам проверки можно сделать вывод, что ряд стационарный.

2. Для прогнозирования разделите временной ряд на обучающую, валидационную и контрольную выборки.

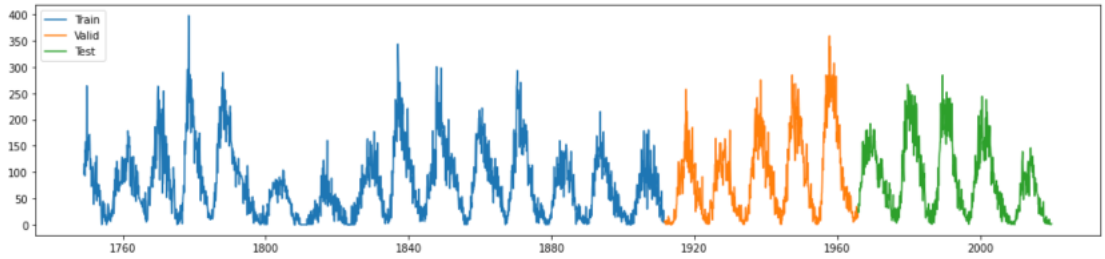

Рисунок 7 – разделение данных на train, valid & test.

Данные были разделены на train, valid и test выборки по принципу 60%/20%/20%. Результат можно увидеть на рисунке 7.

3. Примените модель ARIMA для прогнозирования значений данного временного ряда.

Для прогнозирования была использована модель «SARIMAX» из пакета «statsmodels».

Для ускорения работы модели «SARIMAX», исходные данные были сжаты по годам.

Модель SARIMAX имеет следующий вид - SARIMA(p,d,q)(P,D,Q)S.

Где:

- p — порядок модели AR(p);
- d — порядок интегрирования исходных данных;
- q — порядок модели MA(q);
- P — порядок сезонной составляющей SAR(P);
- D — порядок интегрирования сезонной составляющей;
- Q — порядок сезонной составляющей SMA(Q);
- s — размерность сезонности(месяц, квартал и т.д.).

Модель запускалась со следующими параметрами:

- p — 3;
- d — 0;
- q — 3;
- P — 1;
- D — 1;
- Q — 0;
- s — 11.

Результат работы можно увидеть на рисунке 8. Ошибка rmse составила «727».

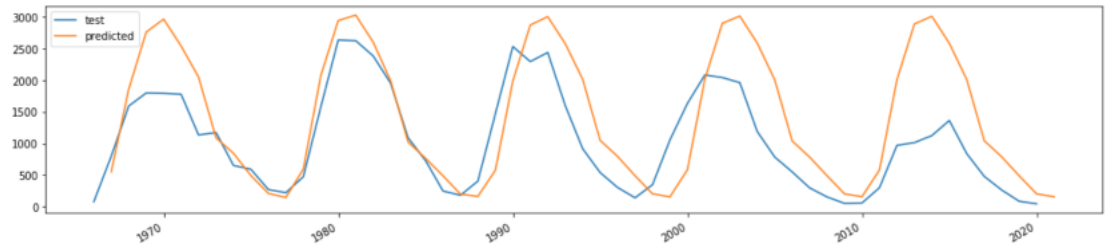

Рисунок 8 – предсказания sarimax модели.

Далее для подбора параметров был подключен пакет «pmdarima». В нем есть удобный механизм «auto_arima» для подбора параметров.

После запуска, «auto_arima» подобрал следующие параметры:

- p — 1;
- d — 0;
- q — 0;
- P — 2;
- D — 1;
- Q — 1;
- s — 11.

Результат работы можно увидеть на рисунке 9. Ошибка rmse составила «492».

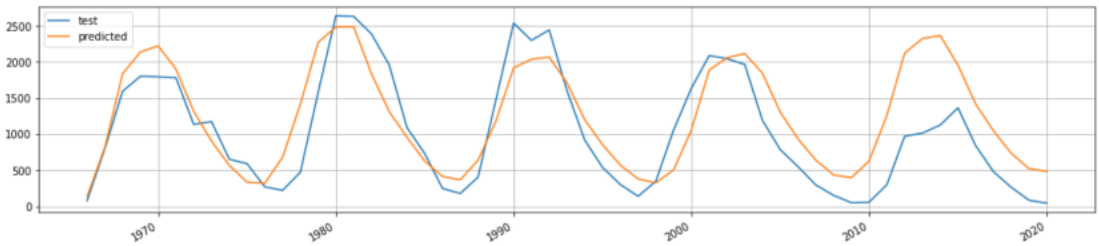

Рисунок 9 – предсказания auto_arima модели.

4. Повторите эксперимент по прогнозированию, реализовав рекуррентную нейронную сеть (с как минимум 2 рекуррентными слоями).

Для решения задачи прогнозирования была выбрана архитектура с 2мя LSTM слоями. Архитектуру можно посмотреть в таблице 1.

Таблица 1 – Архитектура нейронной сети.

Слой		**Размер**	**Активация**
Входной	-	60	-
1	Conv1D(filters=60, kernel =5)		-
2	LSTM(60, return_sequences=True)		tanh
3	LSTM(60, return_sequences=True)		tanh
4	Dropout(0.5)		
5	FC	10	Relu
6	Dropout(0.5)		

Выходной	FC	1	

Тренировка нейросети была запущена со следующими параметрами:

- epochs – 20;
- batch size - 32.

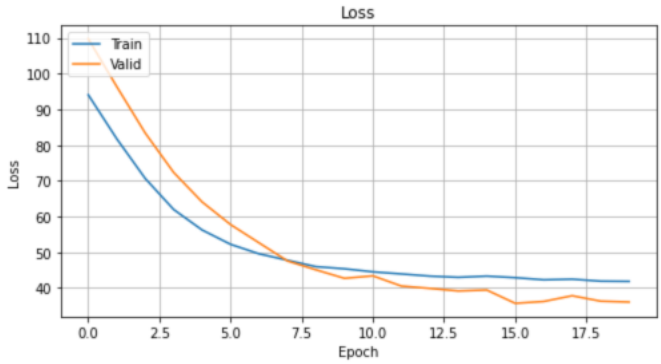

Рисунок 10 – график изменения loss модели.

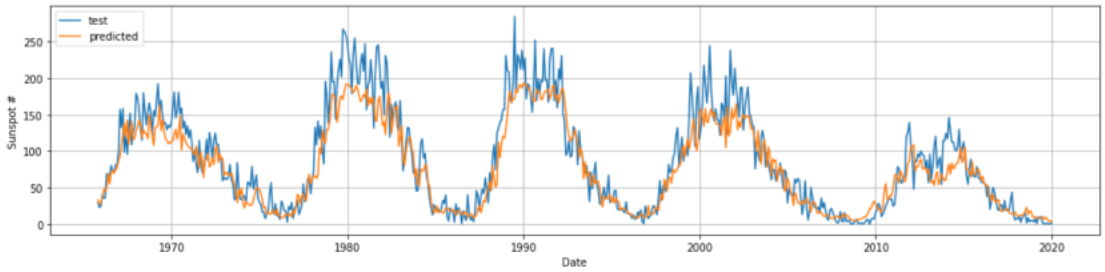

Рисунок 11 – предсказания rnn модели.

Результат работы можно увидеть на рисунке 8. Ошибка rmse составила «29».

5. Сравните качество прогноза моделей.

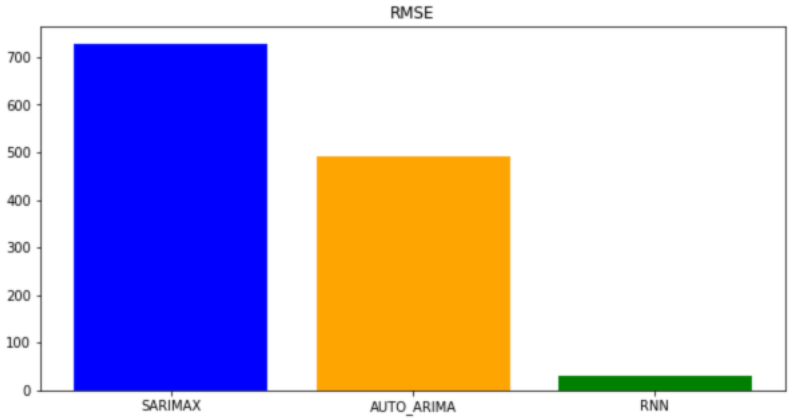

Рисунок 12 – предварительное сравнения качества предсказания моделей.

По рисунку 12 видно, что модель построенная на базе RNN дает значительное преимущество по сравнению с моделями построенными на

SARIMAX. Но это не совсем так. Причина такой большой разницы заключается в том, что готовой модели «SARIMAX» не хватает памяти (google colab - 25gb) для подсчетов (коэффициент сезонности 132). Поэтому для моделей на базе SARIMAX» данные пришлось «сгруппировать» по годам. Отсюда такая большая разница.

При этом, хочется отметить, что «SARIMAX» обучается очень долго. Нейронная сеть обучается значительно быстрее.

В следующем эксперименте, данные были сгруппированы по кварталам (коэффициент сезонности 43), и следующий эксперимент уже более точно показывает разницу между моделями.

Поквартальные данные модели «SARIMAX» показали ошибку rmse «178». На рисунке 13 можно увидеть сравнение предсказанных данных с оригинальными.

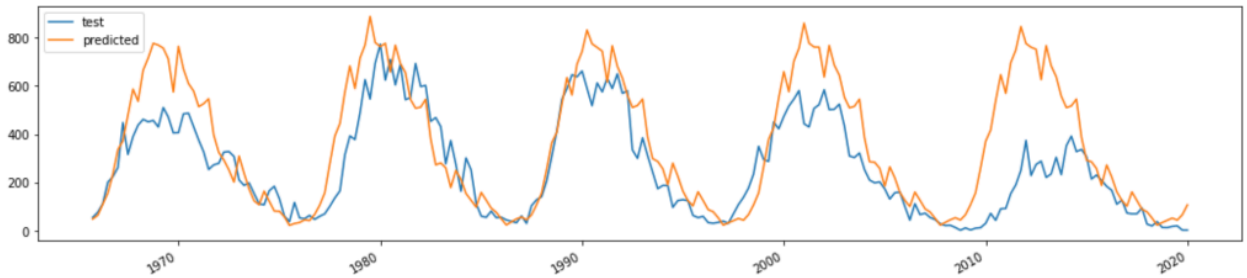

Рисунок 13 – предсказания «SARIMAX» модели на поквартальных данных.

На тех-же поквартальных данных была обучена RNN сеть. Ошибка rmse составила «120». На рисунке 14 можно увидеть сравнение предсказанных данных с оригинальными.

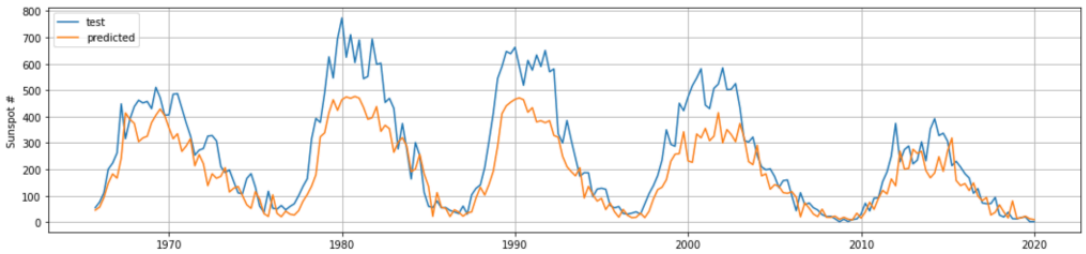

Рисунок 14 – предсказания «RNN» модели на поквартальных данных.

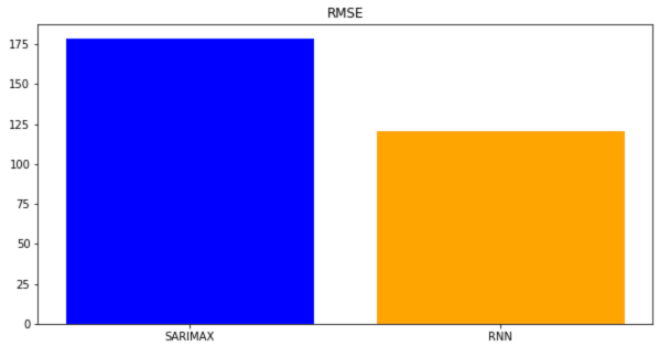

Рисунок 15 – сравнения качества предсказания моделей на поквартальных данных.

По рисунку 15 видно, что для данного набора данных модель на базе RNN показала более точные результаты чем «SARIMAX», при этом, обучение RNN модели заняло гораздо меньше времени, чем обучение модели «SARIMAX».

Вывод.

В ходе выполнения лабораторной работы я провел анализ временного ряда «среднемесячное число пятен на солнце», вычислил основные характеристики и отобразил в виде графиков. Так же я построил 2 модели для прогнозирования. Первая модель была на базе «SARIMAX», вторая на базе рекуррентных нейронных сетей. Обе модели показали хороший результат в прогнозировании, но на конкретном наборе данных точность у RNN была выше.