# ИССЛЕДОВАНИЕ СТАТИСТИЧЕСКИХ МЕТОДОВ ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ С ТРЕНДОМ И СЕЗОННОСТЬЮ

STUDY OF STATISTICAL METHODS FOR FORECASTING TIME SERIES WITH TREND AND SEASONALITY

**Дауб Игорь Сергеевич,** Магистрант (обучается на магистра, есть степень бакалавра) СамГУПС Прикладная математика, информатика и информационные системы"

**Daub Igor Sergeevich,** Master's student (studying for a master's degree, there is a bachelor's degree) SamGUPS Applied Mathematics, Informatics and Information Systems ", Email: igordaub@gmail.com

**Аннотация**

Целью данной статьи является обзор трех статистических моделей прогнозирования временных рядов - Экспоненциальное сглаживание Хольта-Винтерса, SARIMA, Prophet и проведение экспериментов на нескольких временных рядах с месячными данными. Временные ряды с такой разбивкой будут рассмотрены исходя из того, что они больше всего подходят для прогнозирования рассматриваемыми способами в виду отсутствия в них комплексных сезонностей, с которыми может работать только Prophet из рассматриваемых в данной статье моделей прогнозирования.

**Annotation**

The purpose of this article is to review three statistical models for time series forecasting - Holt-Winters Exponential Smoothing, SARIMA, Prophet and performing experiments on multiple time series with monthly data. Time series with such a breakdown will be considered based on the fact that they are most suitable for

forecasting by the methods under consideration due to the absence of complex seasonality in them, with which only Prophet from the forecasting models considered in this article can work.

**Ключевые слова:** статистические модели временных рядов, прогнозирование, Винтерса, SARIMA, Prophet.

**Key words:** statistical time series models, forecasting, Winters, SARIMA, Prophet.

В каждом временном ряду будут присутствовать тренд и сезонность, потому в данной статье менее комплексные модели – наивная, скользящее среднее, экспоненциальное сглаживание, двойное экспоненциальное сглаживание и т.д., которые не учитывают либо тренд, либо сезонность, либо тренд и сезонность, рассматриваться не будут.

Эксперимент будет проводиться с помощью языка Python. Для прогнозирования будет использоваться библиотека statsmodels и fbprophet. Операции над данными будут совершаться с помощью pandas, построение графиков с помощью matplotlib. Для вычисления метрик производительности будет использоваться sklearn

1. Используемые временные ряды и их характеристики

Рассмотрим временные ряды, выбранные для эксперимента. Эксперимент будет проведен с использованием трех временных рядов:

- Пройденные мили (https://www.kaggle.com/bulentsiyah/for-simple-exercises-time-series-forecasting?select=Miles_Traveled.csv) – 588 записей
- Продажи пива (источник: https://www.kaggle.com/bulentsiyah/for-simple-exercises-time-series-forecasting?select=BeerWineLiquor.csv) – 324 записи
- Посетители аэропорта(https://www.kaggle.com/andreazzini/international-airline-passengers) - 145 записей.

Поскольку в каждом наборе данных вычисления представлены по месяцам, то валидационная выборка у каждого ряда будет составлять последние 24 записи.

1.1. График и декомпозиция временного ряда «Пройденные мили»

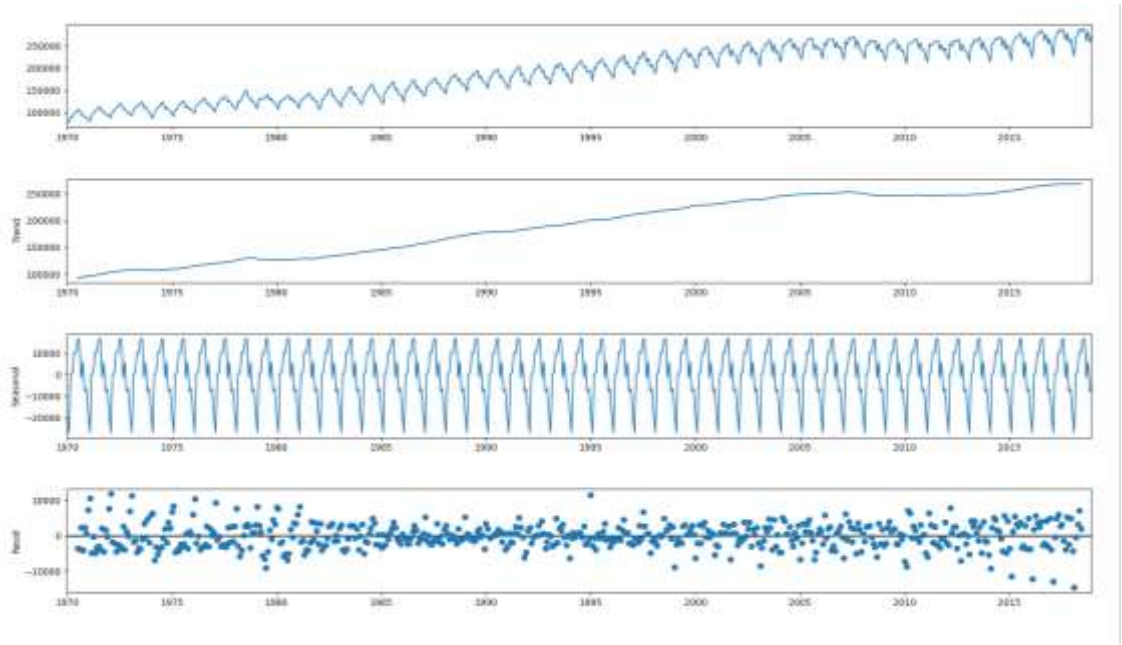

Рисунок 1 – Декомпозиция временного ряда «Пройденные мили»

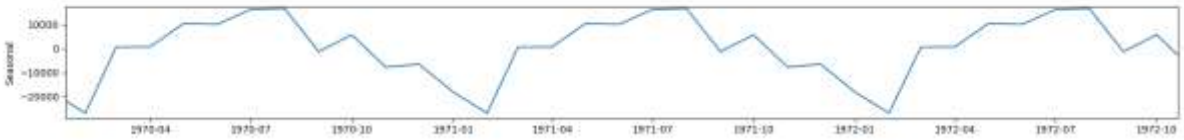

Рисунок 2 – Увеличенный график сезонности временного ряда «Пройденные мили»

В данном временном ряде присутствуют тренд и сезонность. Сезонность и тренд аддитивны. Как видно из увеличенного графика декомпозиции сезонности, ее период составляет 12 измерений.

1.2. График и декомпозиция временного ряда «Продажи пива»

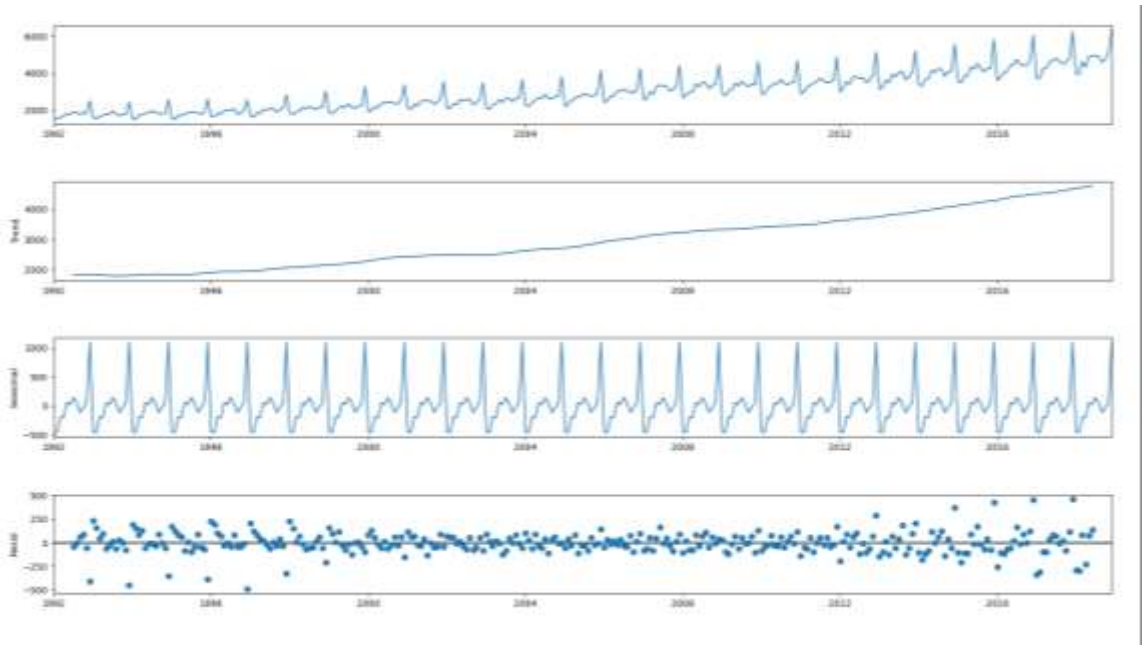

Рисунок 3 – Декомпозиция временного ряда «Продажи пива»

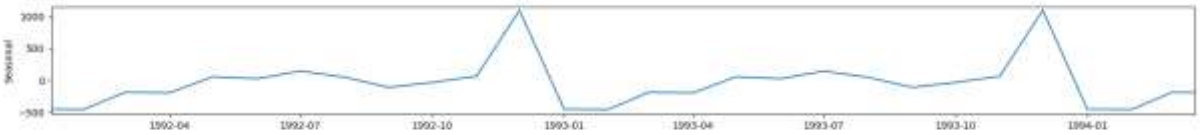


Рисунок 4 – Увеличенный график сезонности временного ряда «Продажи пива»

В данном временном ряду присутствуют тренд и сезонность. Сезонность мультипликативна, тренд аддитивен. Как видно из увеличенного графика декомпозиции сезонности, ее период составляет 12 измерений.

1.3. График и декомпозиция временного ряда «Посетители аэропорта»

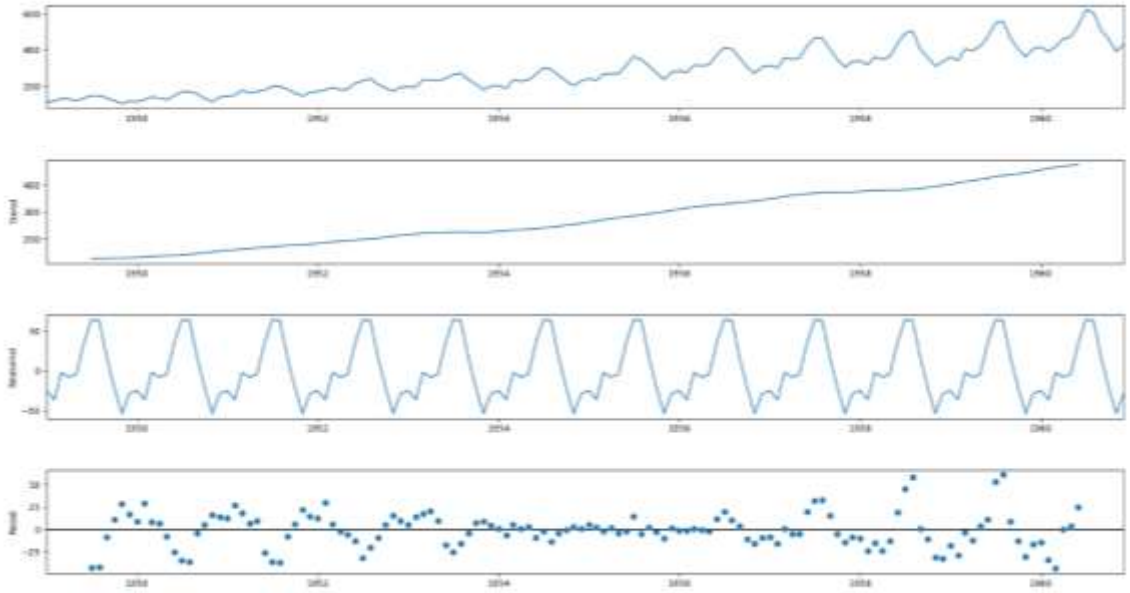

Рисунок 5 – Декомпозиция временного ряда «Посетители аэропорта»

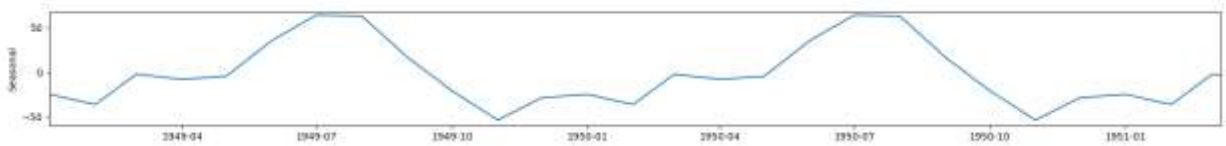

Рисунок 6 – Увеличенный график сезонности временного ряда «Посетители аэропорта»

В данном временном ряде присутствуют тренд и сезонность. Сезонность и тренд мультипликативны. Как видно из увеличенного графика декомпозиции сезонности, ее период составляет 12 измерений.

## 2. Параметры оценки моделей прогнозирования

Для оценки производительности моделей прогнозирования будут использованы следующие метрики:

- MSE (Mean Squared Error) = $\frac{1}{n}\sum(\text{Прогноз} - \text{Действительные значения})^2$
- RMSE (Root Mean Squared Error) = $\sqrt{\text{MSE}}$
- MAE (Mean Absolute Error) = $\frac{1}{n}\sum|\text{Прогноз} - \text{Действительныез начения}|$
- MAPE (Mean Absolute Percentage Error) = $\frac{1}{n}\sum\frac{|\text{Прогноз}-\text{Действительные значения}|}{\text{Действительные значения}}$

В случаях, когда большие ошибки важны и они могут привести к серьезным последствиям, следует уделять больше внимания MSE. В том случае, когда приоритетом обладают малые ошибки, следует уделить внимание MAE.

## 3. Prophet

Prophet – библиотека с открытым кодом для прогнозирования одномерных временных рядов, разработанная Facebook, доступная для языков Python и R. Как заявляют создатели, Prophet оптимизирован для решения задач, которые встречались в самом Facebook со следующими характеристиками:

- часовые, дневные или недельные наблюдения с как минимум несколькими месяцами (лучше год) историческими данными;

- явные множественные «применимые к людям» сезонности: по дням недели, времени года;
- важные праздники, которые происходят на нерегулярных промежутках известны заранее;
- адекватное значение аномалий или пропущенных наблюдений;
- исторические изменения тренда, например, из-за запуска продукта;
- тренды, которые имеют нелинейные кривые роста, когда тренд достигает естественного предела или насыщения. [1]

Prophet по умолчанию делает прогноз автоматически, т.е. не нуждается в подстройке параметров, однако опытный аналитик может улучшить работу библиотеки с помощью ее точной подстройки, например, явного указания трендов, которые были пропущены во время полностью автоматического прогнозирования.

Несмотря на то, что в описании указано наиболее подходящее использование при часовых, дневных или недельных наблюдения, в [2] говорится о том, что месячные наблюдения также можно использовать, однако при использовании месячных наблюдений необходимо делать прогнозы также по месяцам. при запросе прогнозов с иными интервалами измерений, например, дневных прогнозов у модели Prophet, что была настроена по месячным данным, прогнозные значения будут неточными.

### 3.1. Прогнозирование

Для прогнозирования с помощью Prophet, данные необходимо предоставить в определенном формате – первая колонка должна называться «ds» и должна содержать дату и время, вторая колонка должна называться «y» и содержать значения

#### 3.1.1. Продажи пива

Прогнозирование на валидационной выборке:

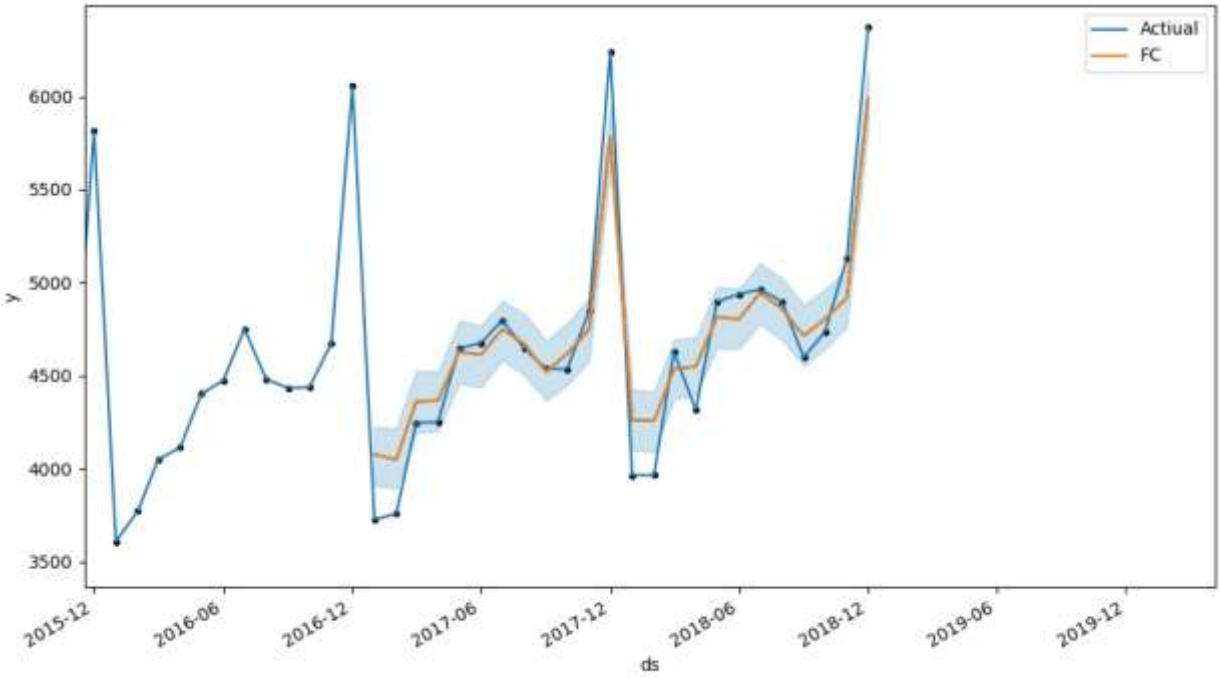


Рисунок 7 – Прогноз временного ряда «Продажи пива» на валидационной выборке с помощью Prophet

Результаты метрик производительности:

- MAE = 151.8522424865687
- MSE = 39036.177265587474
- RMSE = 197.5757507023255
- MAPE = 0.03300558079637045

Прогноз на 48 измерений:

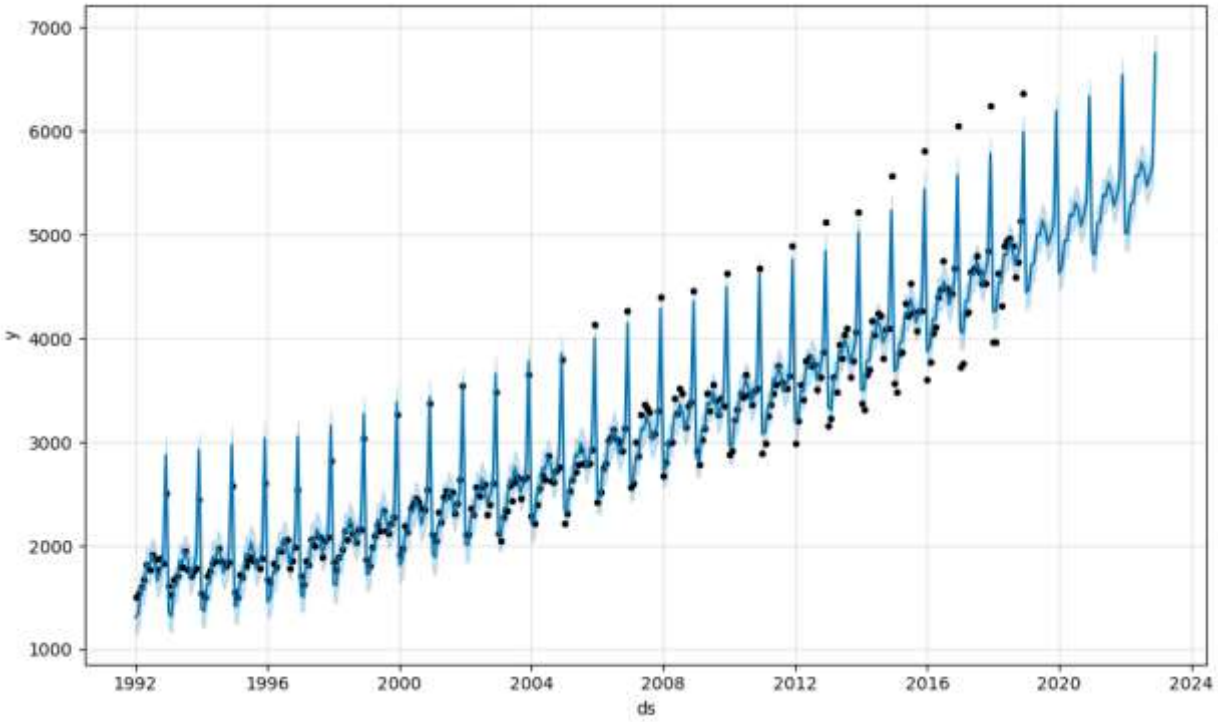


Рисунок 8 – Прогноз временного ряда «Продажи пива» на 48 измерений с помощью Prophet

### 3.1.2. Посетители аэропорта

Прогнозирование на валидационной выборке:

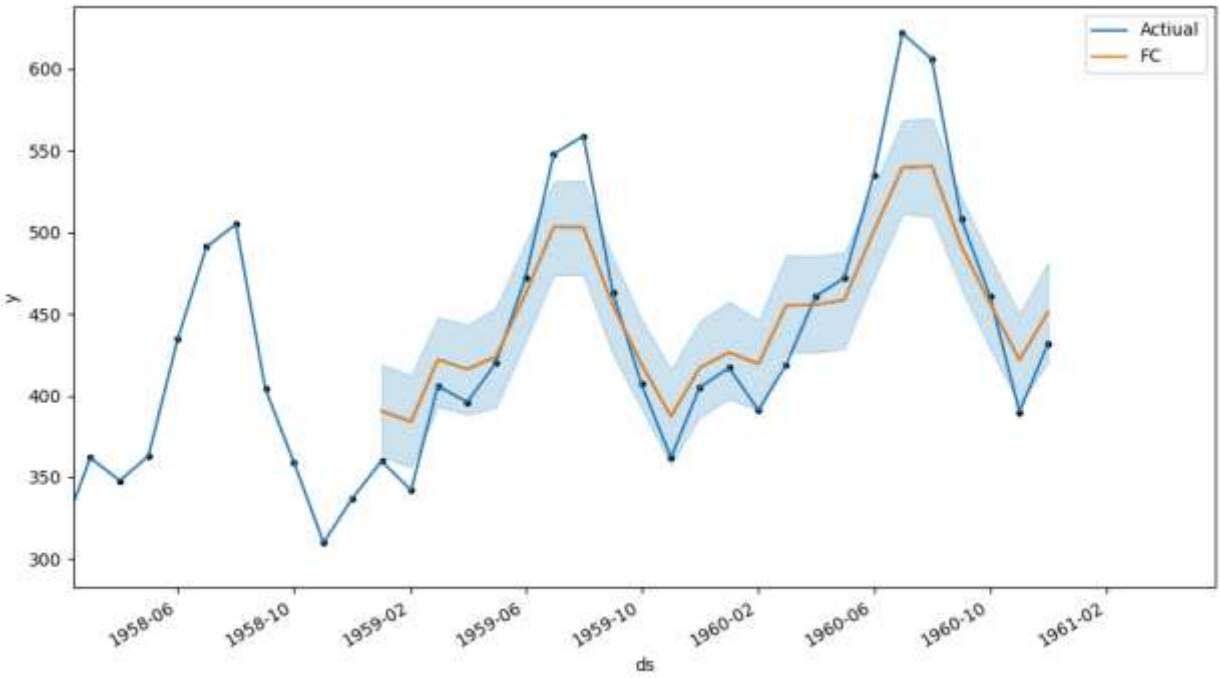


Рисунок 9 – Прогноз временного ряда «Посетители аэропорта» на валидационной выборке с помощью Prophet

Результаты метрик производительности:

- MAE = 26.12854691299727
- MSE = 1080.8563037464849
- RMSE = 32.87637911550609
- MAPE = 0.05625603794576667

Прогноз на 48 измерений:

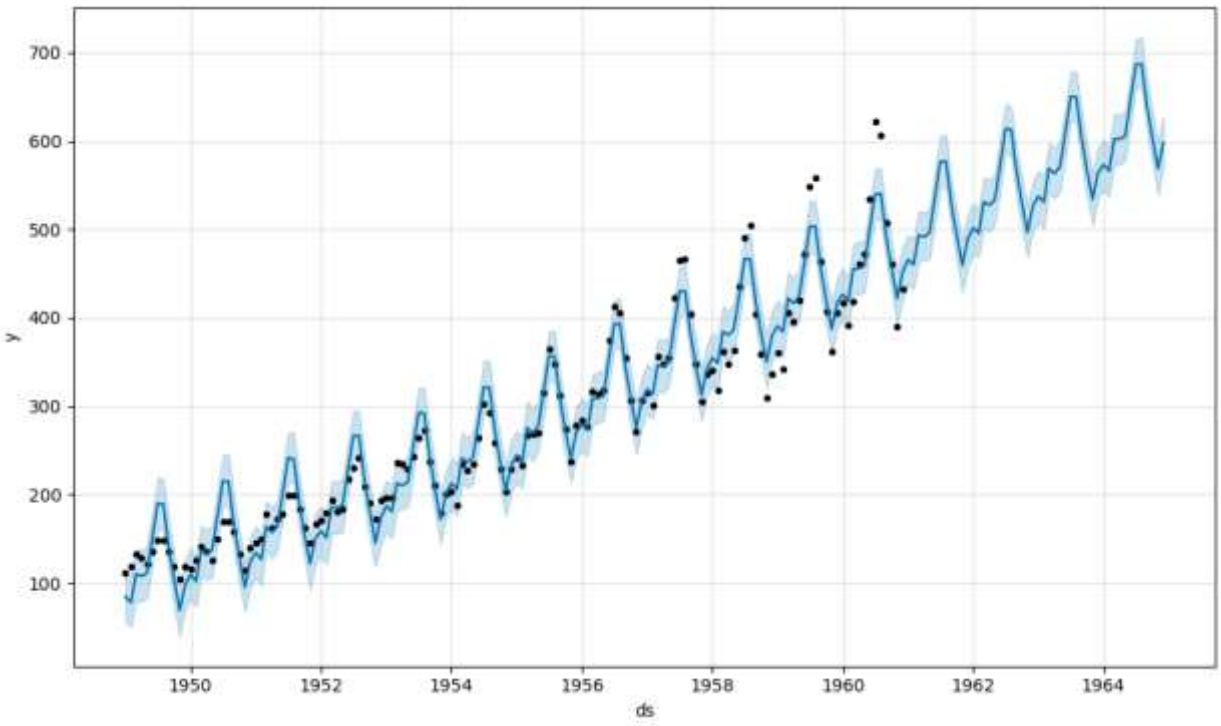


Рисунок 10 – Прогноз временного ряда «Посетители аэропорта» на 48 измерений с помощью Prophet

### 3.1.3. Пройденные мили

Прогнозирование на валидационной выборке:

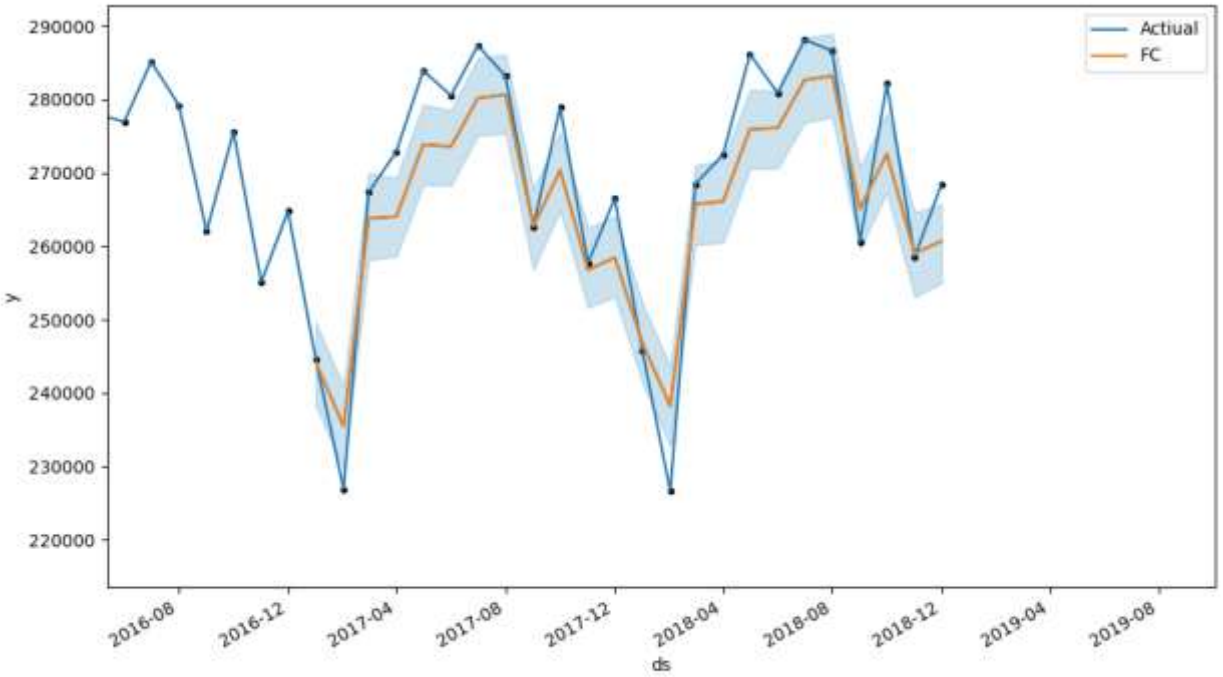


Рисунок 11 – Прогноз временного ряда «Пройденные мили» на валидационной выборке с помощью Prophet

Результаты метрик производительности:

- MAE = 5563.585846730039
- MSE = 43097217.32926529
- RMSE = 6564.847091080285
- MAPE = 0.020755049627825525

Прогноз на 48 измерений:

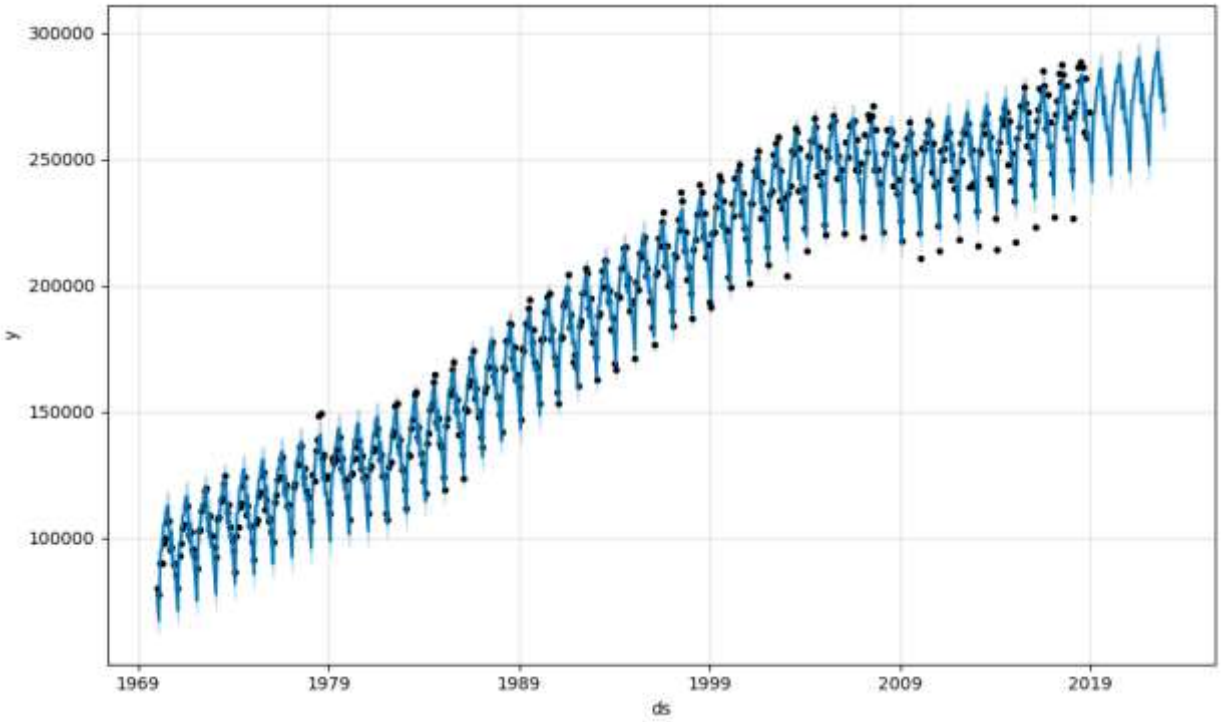


Рисунок 12 – Прогноз временного ряда «Пройденные мили» на 48 измерений с помощью Prophet

4. Экспоненциальное сглаживание Хольта-Винтерса

Модель Хольта-Винтерса является третьим этапом развития модели экспоненциального сглаживания.

Первым этапом является простое экспоненциальное сглаживание. Данная модель вычисляет прогнозные значения, вычисляя взвешенное среднее, в котором весовой коэффициент, применяемый к значению ряда уменьшается экспоненциально, по направлению к началу временного ряда. Даная модель не учитывает сезонность и тренд.

$$\hat{y}_x = \alpha * y_x + (1 - \alpha) * \hat{y}_{x-1}$$

,где $\hat{y}_x$ – прогнозное значение, $y_x$ – значение временного ряда в точке x. первое прогнозное значение инициализируется первым значением временного ряда, где $\alpha$ - коэффициент сглаживания, 0<α<1

Второй этап – метод Хольта, добавляет к экспоненциальному сглаживанию сглаживание тренда, что позволяет данной модели учитывать тренд, но не сезонность.

$$\ell_x = \alpha * y_x + (1 - \alpha) * \ell_{x-1} + b_{x-1}$$

$$b_x = \beta(\ell_x - \ell_{x-1}) + (1-\beta) * b_{x-1}$$

$$\hat{y}_{x+1} = \ell_x + b_x$$

В первом выражении вместо $\hat{y}_x$ используется $\ell_x$. В данных выражениях $\alpha$ - коэффициент сглаживания уровня, $\beta$ - коэффициент сглаживания тренда, $\ell_x$ – уровень или базовая линия, $b_x$ - тренд

Третий этап – метод Хольта-Винтерса. Данная модель учитывает и тренд, и сезонность.

Для мультипликативной сезонности используется следующая формула:

$$\ell_x = \alpha * \frac{y_t}{s_{x-m}} + (1-\alpha)(\ell_{x-1} + b_{x-1})$$

$$b_x = \beta(\ell_x - \ell_{x-1}) + (1-\beta) * b_{x-1}$$

$$s_x = \gamma \frac{y_x}{\ell_{x-1} + b_{x-1}} + (1-\gamma) * s_{x-m}$$

$$\hat{y}_{x+h|x} = (\ell_x + b_x * h) * s_{x-m+h_m^+}$$

где m – сезонные период, $h_m^+ = \lfloor (\mathrm{p}-1)\ mod\ m \rfloor + 1$, $\alpha, \beta, \gamma$ находятся в районе от 0 до 1, $\ell_x$ – уровень или базовая линия, $b_x$ – тренд, $s_x$-сезонность.

Аддитивная сезонность используется следующая формула:

$$\ell_x = \alpha * (y_x - s_{x-m}) + (1-\alpha)(\ell_{x-1} + b_{x-1})$$

$$b_x = \beta(\ell_x - \ell_{x-1}) + (1-\beta) * b_{x-1}$$

$$s_x = \gamma(y_x - \ell_{x-1} - b_{x-1}) + (1-\gamma) * s_{x-m}$$

$$\hat{y}_{x+h|x} = \ell_x + b_x * h + s_{x-m+h_m^+}$$

, где $\ell_x$ – уровень или базовая линия, $b_x$ – тренд, $s_x$-сезонность, m – сезонный период, $h_m^+ = \lfloor (\mathrm{p}-1)\ mod\ m \rfloor + 1$, $\alpha, \beta, \gamma$ находятся в районе от 0 до 1

## 4.1. Прогнозирование

### 4.1.1. Продажи пива

Как было описано выше, данный ряд имеет аддитивный тренд, мультипликативную сезонность и период сезонности равный 12. Указываем данный параметры в коде и выполняем прогнозирование на валидационной выборке:

fitted_model = ExponentialSmoothing(training_period['beer'], trend='add', seasonal='mul', seasonal_periods=12).fit()

validation_predictions = fitted_model.forecast(len(validation_period)).

Графическое отображение результатов:

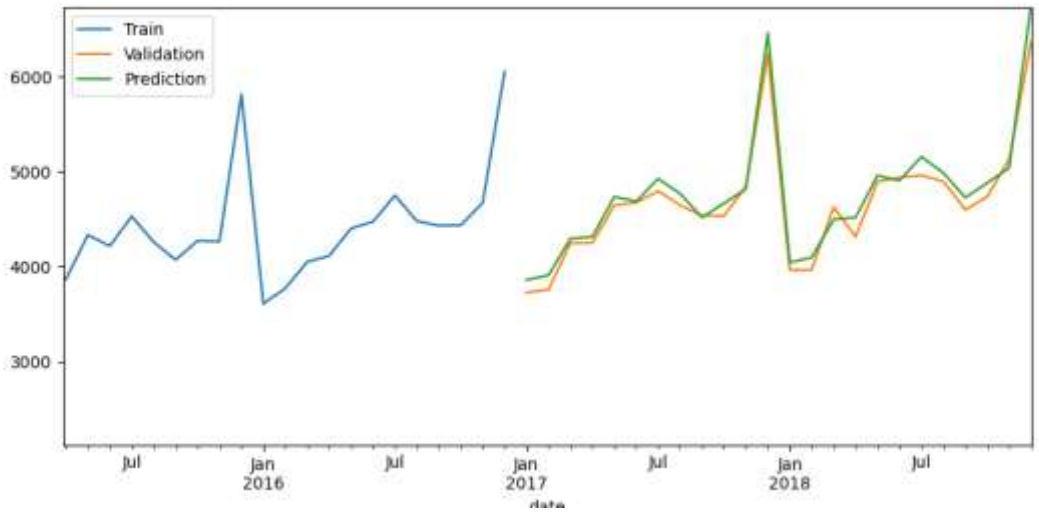


Рисунок 13 – Прогноз временного ряда «Продажи пива» на валидационной выборке с помощью экспоненциального сглаживания Хольта-Винтерса

Вычисляем метрики производительности:

- MAE = 117.6732963899102
- MSE = 20135.635266858462
- RMSE = 141.9000890304811
- MAPE = 0.02475035746209278

Выполняем прогноз на 48 измерений вперед:

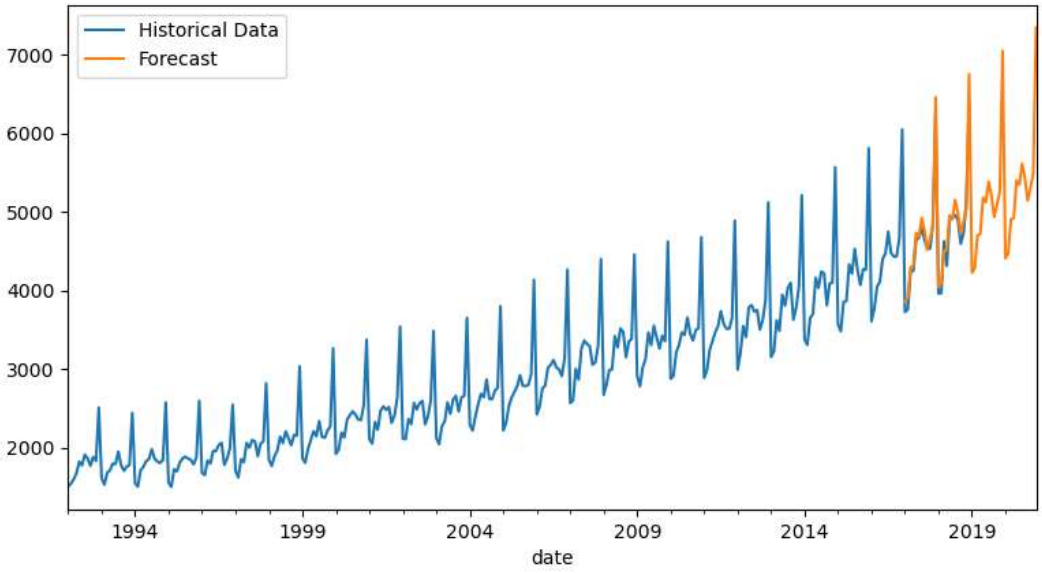


Рисунок 14 – Прогноз временного ряда «Продажи пива» на 48 измерений с помощью экспоненциального сглаживания Хольта-Винтерса

### 4.1.2 Посетители аэропорта

Как было описано выше, данный ряд имеет мультипликативные тренд и сезонность. Период равен 12. Указываем данный параметры в коде и выполняем прогнозирование на валидационной выборке:

```
fitted_model = ExponentialSmoothing(training_period['passengers'], trend='mul', seasonal='mul', seasonal_periods=12).fit()
test_predictions = fitted_model.forecast(len(validation_period))
```

Графическое отображение результатов:

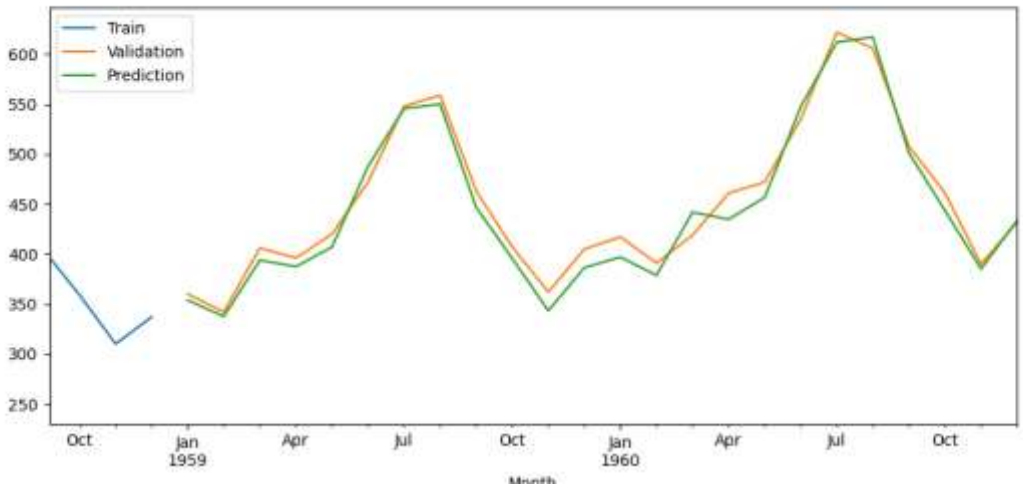


Рисунок 15 – Прогноз временного ряда «Посетители аэропорта» на валидационной выборке с помощью экспоненциального сглаживания Хольта-Винтерса

Вычисляем метрики производительности:

- MAE = 12.379391964591376
- MSE = 192.57742277726808
- RMSE = 13.87722676824401
- MAPE = 0.028128679964816414

Выполняем прогноз на 48 измерений вперед:

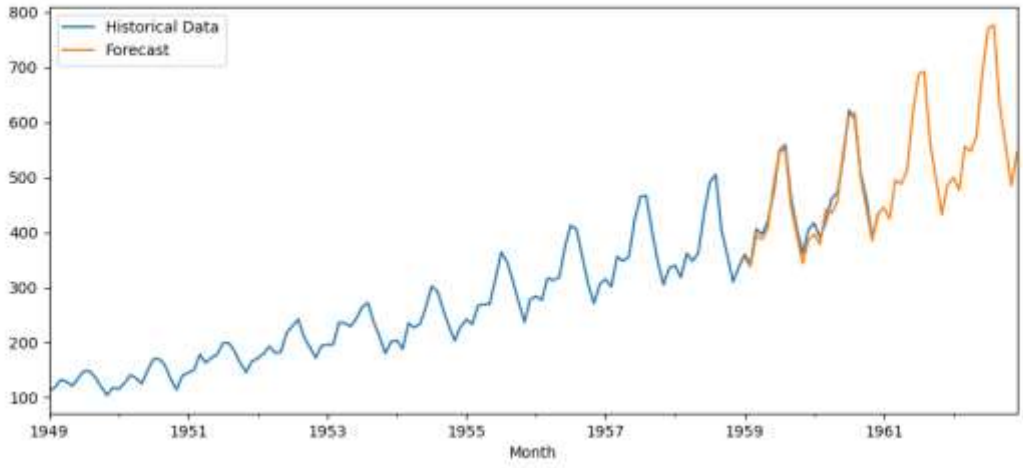


Рисунок 16 – Прогноз временного ряда «Посетители аэропорта» на 48 измерений с помощью экспоненциального сглаживания Хольта-Винтерса

### 4.1.3 Пройденные мили

Как было описано выше, данный ряд имеет аддитивный тренд, мультипликативную сезонность и период сезонности равный 12. Указываем данный параметры в коде и выполняем прогнозирование на валидационной выборке:

```
fitted_model = ExponentialSmoothing(training_period['TRFVOLUSM227NFWA'], trend='add', seasonal='mul',seasonal_periods=12).fit()
test_predictions = fitted_model.forecast(len(validation_period))
```

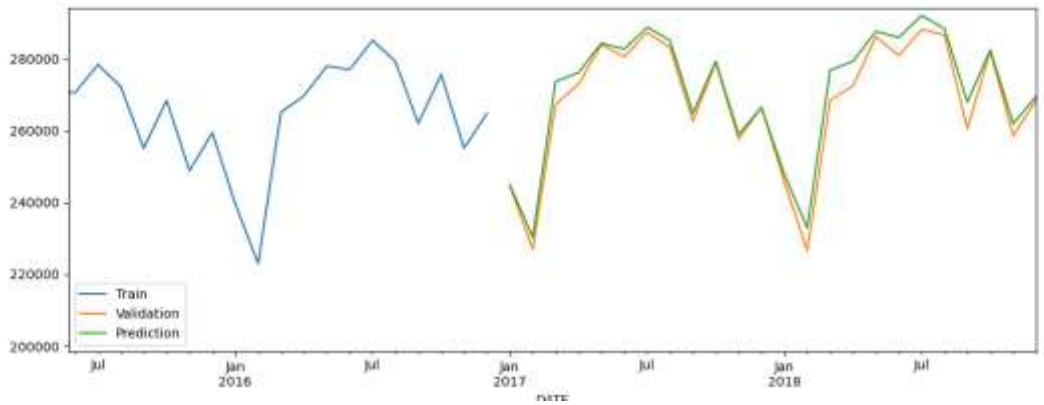


Рисунок 17 – Прогноз временного ряда «Пройденные мили» на валидационной выборке с помощью экспоненциального сглаживания Хольта-Винтерса

Вычисляем метрики производительности:

- MAE = 2947.49522626071
- MSE = 14621080.796941007
- RMSE = 3823.752188223108
- MAPE = 0.011179349177374507

Выполняем прогноз на 48 измерений вперед:

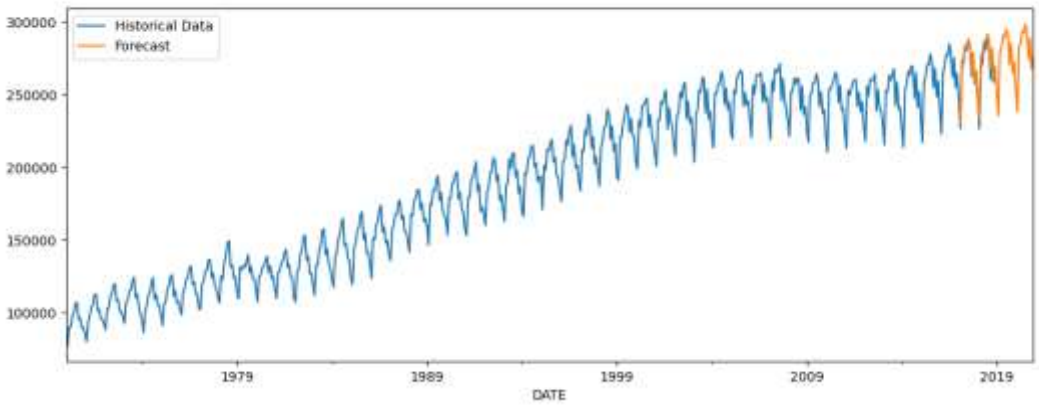


Рисунок 18 – Прогноз временного ряда «Пройденные мили» на 48 измерений с помощью экспоненциального сглаживания Хольта-Винтерса

## 5 SARIMA

Авторегрессионное интегрированное скользящее среднее является обобщением модели авторегрессионного скользящего среднего. Эти модели используются при работе с временными рядами для более глубокого понимания данных или предсказания будущих точек ряда.

Эта аббревиатура носит описательный характер и отражает ключевые аспекты самой модели. Trend Elements:

- p: Порядок авторегрессии.
- d: Порядок дифференцирования.
- q: Порядок скользящего среднего.

Такая модель учитывает тренд, но не учитывает сезонность. при работе с временными рядами, обладающими сезонностью, используют сезонную ARIMA – SARIMA(p,d,q)(P,D,Q)s.В данном случае (p,d,q) – несезонные параметры, а (P,D,Q) являются тем же самым, но применяются к сезонному компоненту. s-период.

Из-за большого количества гиперпараметров, для подстройки модели будет использован поиск по сетке(Grid-Search). Суть данного метода заключается в переборе всех параметров из определенной указанной области с подстройкой модели под каждый, а затем путем сравнения ошибки или других параметров будет выбран оптимальный набор гиперпараметров.

Для GridSearch в данной статье используется сравнение моделей по AIC (Akaike Information Critera). AIC – параметр, который характеризует точность подстройки модели и простоту модели, преобразую в число.

### 5.1 Прогнозирование

#### 5.1.2 Пройденные мили

Лучшие параметры для данного временного ряда, исходя из используемого Grid search - SARIMA(1, 1, 1)x(1, 1, 1, 12) - AIC:10551.881340303584.

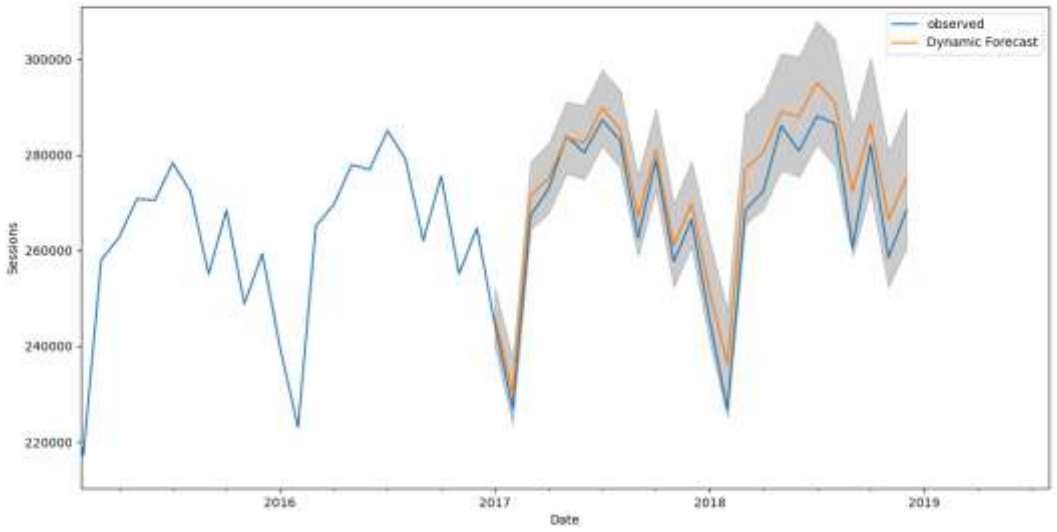


Рисунок 19 – Прогноз временного ряда «Пройденные мили» на валидационной выборке с помощью SARIMA

Значения метрик производительности:

- MAE = 4844.038751652686
- MSE = 31611148.289329242
- RMSE = 5622.379237416242
- MAPE = 0.01834451636094103

Прогноз на 48 измерений вперед:

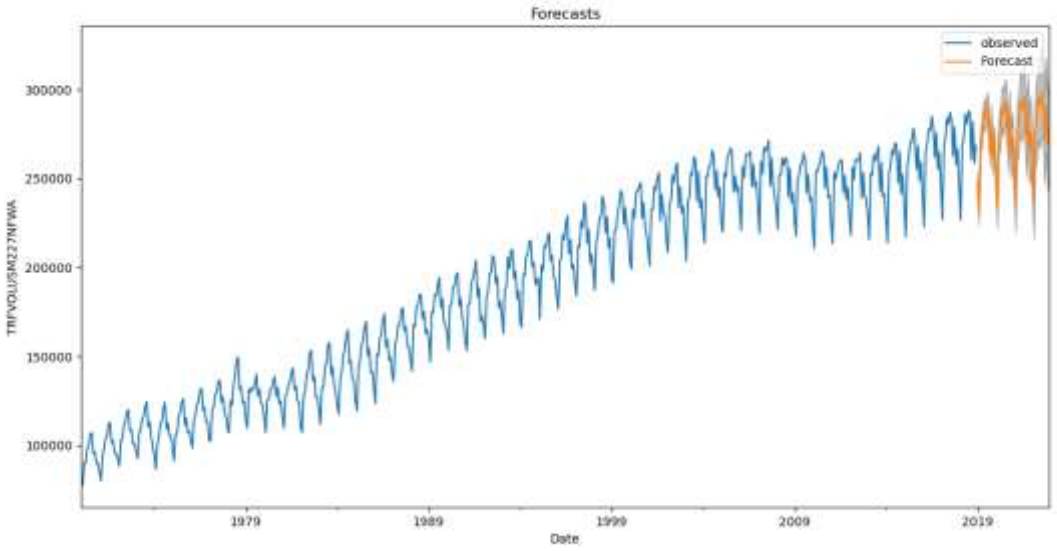


Рисунок 20 – Прогноз временного ряда «Пройденные мили» на 48 измерений с помощью SARIMA

### 5.1.3 Посетители аэропорта

Лучшие параметры для данного временного ряда, исходя из используемого Grid search - SARIMA(0, 1, 1)x(1, 1, 1, 12) - AIC:920.3192974989164

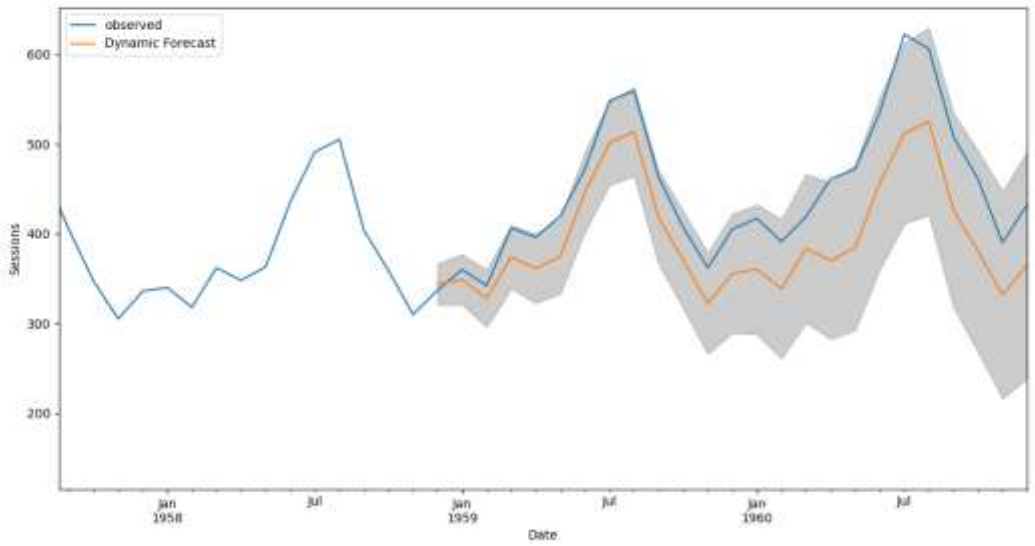


Рисунок 21 – Прогноз временного ряда «Посетители аэропорта» на валидационной выборке с помощью SARIMA

Значения метрик производительности:

- MAE = 52.483505251908554
- MSE = 3439.0564221990317

- RMSE = 58.6434687087917
- MAPE = 0.11357481366300912

Прогноз на 48 измерений вперед:

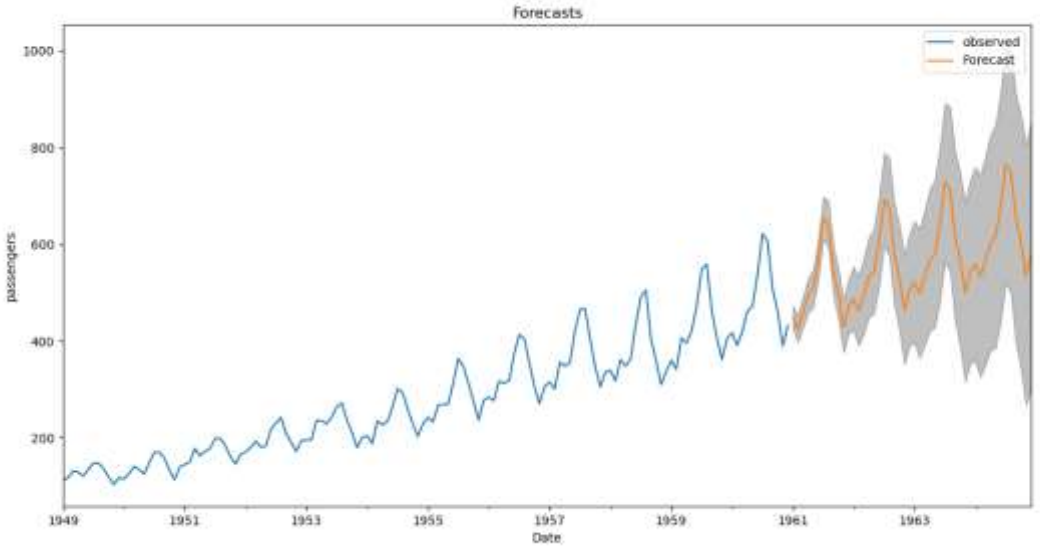


Рисунок 22 – Прогноз временного ряда «Посетители аэропорта» на 48 измерений с помощью SARIMA

### 5.1.4 Продажи пива

Лучшие параметры для данного временного ряда, исходя из используемого Grid search - SARIMA(1, 1, 1)x(1, 1, 1, 12) - AIC:3472.130809344972

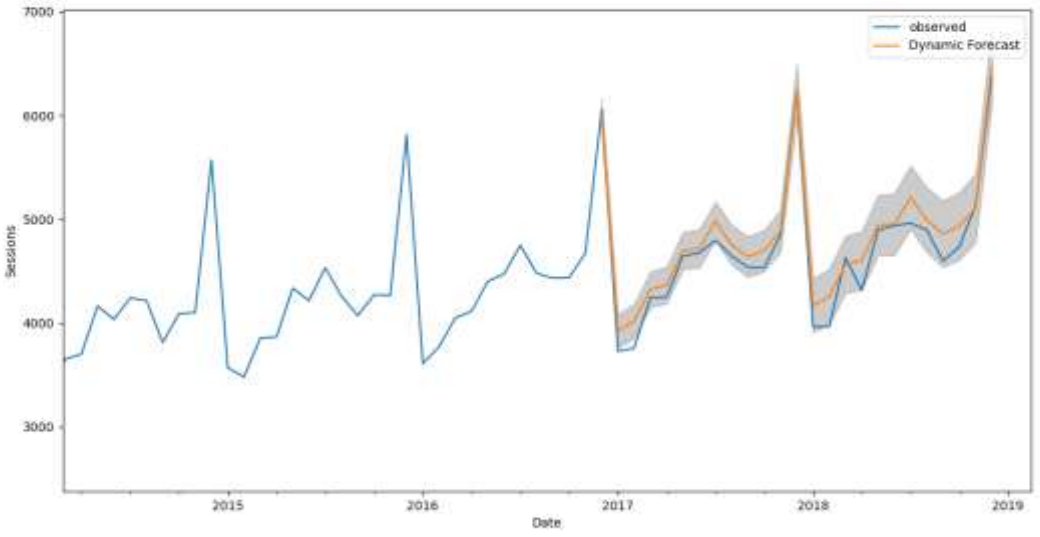


Рисунок 23 – Прогноз временного ряда «Продажи пива» на валидационной выборке с помощью SARIMA

Значения метрик производительности:

- MAE = 128.6974787910323
- MSE = 24772.021397222514
- RMSE = 157.39130025901213
- MAPE = 0.029133116176326194

Прогноз на 48 измерений вперед

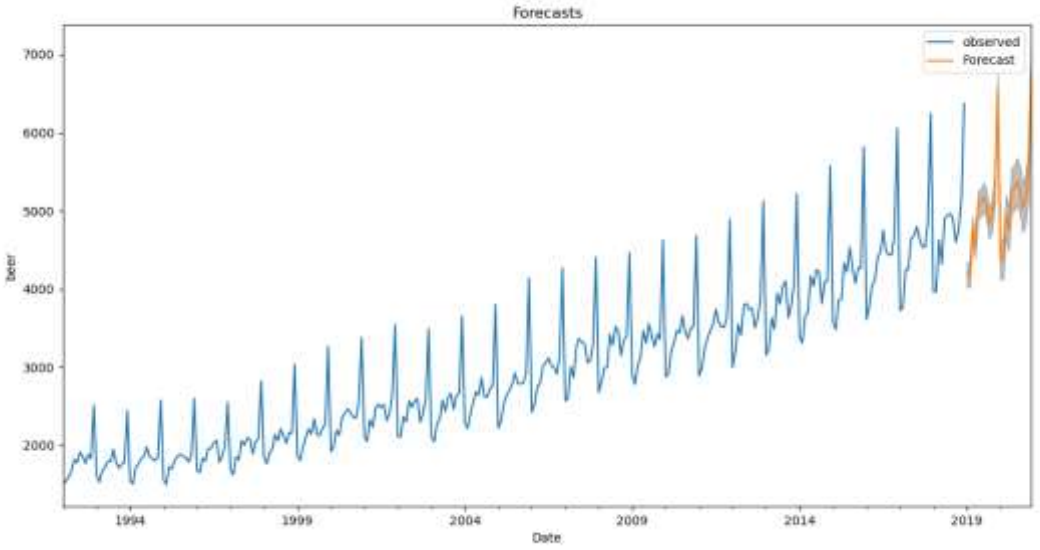


Рисунок 24 – Прогноз временного ряда «Продажи пива» на 48 измерений с помощью SARIMA

## 6 Анализ результатов

Как можно заметить из метрик производительности, метод Хольта-Винтерса показал лучшие результаты на каждом временном ряду.

Таблица 1 – Значения метрик производительности для временного ряда «Продажи пива»

| **«Продажи пива»** | | | |
|---|---|---|---|
| Метрики | SARIMA | Holt-Winters | Prophet |
| MAE | 128.6974787910323 | 117.6732963899102 | 151.8522424865687 |
| MSE | 24772.021397222514 | 20135.635266858462 | 39036.177265587474 |
| RMSE | 157.39130025901213 | 141.9000890304811 | 197.5757507023255 |
| MAPE | 0.029133116176326194 | 0.02475035746209278 | 0.03300558079637045 |

Таблица 2 – Значения метрик производительности для временного ряда «Посетители аэропорта»

| **«Посетители аэропорта»** | | | |
|---|---|---|---|
| Метрики | SARIMA | Holt-Winters | Prophet |
| MAE | 52.483505251908554 | 12.379391964591376 | 26.12854691299727 |
| MSE | 3439.0564221990317 | 192.57742277726808 | 1080.8563037464849 |
| RMSE | 58.6434687087917 | 13.87722676824401 | 32.87637911550609 |
| MAPE | 0.11357481366300912 | 0.028128679964816414 | 0.05625603794576667 |

Таблица 3 – Значения метрик производительности для временного ряда «Пройденные мили»

| **«Пройденные мили»** | | | |
|---|---|---|---|
| Метрики | SARIMA | Holt-Winters | Prophet |
| MAE | 4844.038751652686 | 2947.49522626071 | 5563.585846730039 |
| MSE | 31611148.289329242 | 14621080.796941007 | 43097217.32926529 |
| RMSE | 5622.379237416242 | 3823.752188223108 | 6564.847091080285 |
| MAPE | 0.01834451636094103 | 0.011179349177374507 | 0.020755049627825525 |

При увеличении валидационной выборки для каждого временного ряда на 12 измерений, были получены следующие результаты.

Таблица 4 – Значения метрик производительности для временного ряда «Продажи пива» на увеличенной валидационной выборке

| **«Продажи пива»** | | | |
|---|---|---|---|
| Метрики | SARIMA | Holt-Winters | Prophet |
| MAE | 102.20287990884626 | 80.49262592832102 | 145.41999918029575 |
| MSE | 16478.0295819655 | 12184.21950844156 | 37245.3787538003 |
| RMSE | 128.36677756322118 | 110.38215212814778 | 192.99061830514017 |
| MAPE | 0.02185450783123929 | 0.017395777665762314 | 0.03170988584258564 |

Таблица 5 – Значения метрик производительности для временного ряда «Посетители аэропорта» на увеличенной валидационной выборке

| **«Посетители аэропорта»** |
|---|

| Метрики | SARIMA | Holt-Winters | Prophet |
|---|---|---|---|
| MAE | 20.881195556859726 | 63.410814051637146 | 26.190382263413724 |
| MSE | 592.6643055877291 | 4868.115610955749 | 986.3184763618891 |
| RMSE | 24.344697689388735 | 69.77188266741659 | 31.405707703567025 |
| MAPE | 0.05214574304876929 | 0.15136704464542663 | 0.061439031258299415 |

Таблица 6 – Значения метрик производительности для временного ряда «Пройденные мили» на увеличенной валидационной выборке

| **«Пройденные мили»** | | | |
|---|---|---|---|
| Метрики | SARIMA | Holt-Winters | Prophet |
| MAE | 2290.6495836541308 | 26745.925364116785 | 5491.942104723106 |
| MSE | 9718122.561838293 | 1040031076.3467519 | 41711782.0467426 |
| RMSE | 3117.3903447977596 | 32249.51280789761 | 6458.465920537369 |
| MAPE | 0.008796488320718186 | 0.10605405086728037 | 0.020662359185322447 |

Как можно заметить, показания метрик производительности с помощью метода Хольта-Винтерса снизились и ARIMA начала показывать лучшие результаты на временных рядах «Пассажиры аэропорта» и «Пройденные мили». Следовательно, однозначного ответа, какая модель прогнозирования лучше, нет. Перед выбором модели для использования на конкретном временном ряду и для определенной задачи, желательно, протестировать все модели, имеющиеся в распоряжении на данном временном ряду, сравнить их показатели и после этого решить, какую выбрать. Несмотря на то, что Prophet в данной статье показал худшие результаты, по сравнению с другими моделями прогнозирования, это не плохая модель прогнозирования. Она рассчитана на более «реалистичные» временные ряды. Однако даже с такими показателями у нее есть преимущества, например, простота использования, заключающаяся в том, что недельную, годовую и дневные сезонности Prophet может определять автоматически и получить прогнозные значения, можно не анализируя временной ряд

## Литература

1. Prophet: forecasting at scale [Электронный ресурс] https://research.fb.com/blog/2017/02/prophet-forecasting-at-scale/ (дата обращения 12.05.2021)
2. Non-Daily Data [Электронный ресурс] https://facebook.github.io/prophet/docs/non-daily_data.html (дата обращения 14.05.2021)