

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**

Кафедра теории функций и стохастического анализа

**ПРИМЕНЕНИЕ МЕТОДОВ ПРОГНОЗИРОВАНИЯ К
АНАЛИЗУ СПРОСА В РОЗНИЧНОЙ ТОРГОВЛЕ**

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 451 группы
направления 38.03.05 — Бизнес-информатика

механико-математического факультета

Ивановой Юлии Игоревны

Научный руководитель

зав.каф., д. ф.-м. н., доцент

С. П. Сидоров

Заведующий кафедрой

д. ф.-м. н., доцент

С. П. Сидоров

Саратов 2025

ВВЕДЕНИЕ

Прогнозирование спроса в розничной торговле критически важно для оптимизации управления запасами и финансового планирования. Точные прогнозы позволяют минимизировать издержки и повышать удовлетворенность клиентов. Современные методы анализа временных рядов (статистические и машинные) предоставляют мощные инструменты для решения этих задач, однако выбор оптимальной модели требует глубокого понимания данных, что подчеркивает актуальность данного исследования.

Цель работы — разработка и сравнение моделей прогнозирования спроса на основе временных рядов, а также создание шаблонного кода для анализа различных временных рядов, включая методы ARIMA и SARIMA.

Задачи исследования:

- Изучить теорию временных рядов и их применение в торговле.
- Провести предварительную обработку данных: устранение выбросов, нормализация, проверка стационарности.
- Построить и сравнить модели для различных данных.
- Разработать шаблонный код для анализа временных рядов.
- Оценить качество моделей с помощью статистических метрик и визуализации.

Структура работы:

1. Предварительные сведения: обозначения, меры разброса, характеристики распределения.
2. Основы временных рядов: тренд, сезонность, цикличность.
3. Методы анализа: ARMA, ARIMA, SARIMA.
4. Статистические методы оценки: анализ ACF и PACF, метод максимального правдоподобия.
5. Практическая часть: применение моделей к данным и их сравнение.

Основное содержание работы

Бакалаврская работа состоит из: введения, трёх теоретических и одной практической главы, заключения, списка использованных источников, приложения.

Введение содержит актуальность темы исследования, цель и задачи исследования.

Первая глава «Предварительные сведения» содержит информацию, необходимую для понимания теоретической информации.

В разделе «Обозначения» представлены ключевые термины и понятия, используемые в работе для анализа временных рядов и прогнозирования спроса в розничной торговле.

Основное внимание уделено:

- Случайным величинам и их классификации (дискретные и непрерывные).
- Характеристикам распределения, включая математическое ожидание, медиану и моду, которые помогают описывать центральные тенденции данных.
- Основам вероятностного анализа, необходимым для интерпретации закономерностей в динамике спроса.

Раздел служит справочной основой для последующего применения статистических методов и моделей прогнозирования.

В разделе «Меры разброса» рассматриваются ключевые статистические показатели, характеризующие степень вариации данных относительно среднего значения: дисперсии - показателю, отражающему средний квадрат отклонений значений случайной величины от её математического ожидания. Этот параметр позволяет количественно оценить разброс данных в квадратных единицах измерения и среднему квадратическому отклонению - величине, представляющей собой квадратный корень из дисперсии. Данный показатель характеризует разброс данных в исходных единицах измерения, что делает его более удобным для практической интерпретации.

Эти метрики играют важную роль в анализе временных рядов, так как позволяют оценить: степень изменчивости спроса, стабильность продаж, надежность прогнозных моделей.

В контексте розничной торговли понимание мер разброса помогает:

- Выявлять периоды повышенной волатильности спроса.
- Оценивать риски управления запасами.
- Принимать обоснованные решения по закупкам и логистике.

В разделе «Характеристики распределения» рассматриваются ключевые характеристики распределения случайной величины, которые помогают в исследовании ее свойств. Основное внимание уделено коэффициенту асимметрии и эксцессу.

Коэффициент асимметрии описывает степень несимметричности распределения. Он позволяет выявить, насколько распределение склоняется влево или вправо от математического ожидания. Отрицательный коэффициент указывает на преобладание отрицательных значений, что делает кривую распределения более пологой слева от среднего значения. Положительный коэффициент, в свою очередь, сигнализирует о большем влиянии положительных отклонений и приводит к более пологой кривой справа.

Коэффициент эксцесса указывает на «остроконечность» или «плосковершинность» формы распределения. Случайные величины с положительным эксцессом имеют более выраженный пик по сравнению с нормальным распределением, тогда как распределения с отрицательным эксцессом выглядят более плоскими. Значения эксцесса могут варьироваться, и близость к -2 может свидетельствовать о формировании более сложной, возможно, двухвершинной структуры кривой.

Второй раздел «Основы временных рядов» состоит из двух разделов: «Определение временного ряда» и «Компоненты временных рядов».

В разделе «Определение временного ряда» рассматривается понятие временного ряда как упорядоченной последовательности данных, фиксируемых в определенные моменты времени. Указано, что данные могут варьироваться от финансовых показателей до метеорологических сведений. Также проводятся различные классификации временных рядов. Выделяются моментные и интервальные ряды, где интервальный ряд учитывает данные, накопленные за определённый временной период, а моментный фиксирует уровень показателя в конкретный момент. Дополнительно рассматриваются классификации по форме представления уровней, парасостоянию между да-

тами, а также по содержанию показателей, где частные и агрегированные показатели обеспечивают разные уровни анализа.

В разделе «Компоненты временного ряда» рассматриваются ключевые элементы, которые помогают понять структуру и динамику временных рядов, а именно: тренд, сезонность, цикличность и случайные колебания.

Тренд представляет собой общее направление изменения наблюдаемого показателя на протяжении времени и может быть выражён в виде функции от времени:

$$U_t = f(t)$$

где U_t — значение временного ряда в момент времени t . Тренды могут быть линейными или нелинейными. Для выделения тренда из временного ряда используются методы как графического анализа, так и количественного, включая метод наименьших квадратов.

Сезонность обозначает регулярные колебания, наблюдаемые в определённые периоды времени. Временной ряд может быть смоделирован с использованием аддитивных или мультипликативных моделей. В аддитивной модели уровень временного ряда выражается как сумма компонент:

$$Y_t = U_t + V_t + E_t$$

где Y_t — наблюдаемое значение, U_t — тренд, V_t — сезонный эффект, а E_t — случайные колебания. В мультипликативной модели компоненты перемножаются:

$$Y_t = U_t \cdot V_t \cdot E_t$$

что предполагает, что сезонные изменения становятся более выраженными с увеличением уровня тренда.

Циклическая компонента временного ряда отражает долгосрочные изменения, связанные с деловыми циклами в экономике, которые не имеют фиксированной периодичности. Для выделения циклической компоненты могут использоваться статистические методы, такие как фильтрация Ходрика-Прескотта.

Случайные колебания возникают под воздействием непредсказуемых

событий, таких как природные катаклизмы или изменения в потребительских предпочтениях. Их характеристика не поддается простому математическому описанию, и их влияние часто рассматривается как шум в анализе временных рядов.

Третий раздел «Методы анализа временных рядов» включает следующие разделы: «Методы скользящего среднего», «Экспоненциальное сглаживание», «ARMA модели», «ARIMA модели», «SARIMA модели».

Методы сглаживания временных рядов делятся на две категории: механическое и аналитическое сглаживание. Механическое сглаживание на основе фактических значений позволяет создать новые уровни временного ряда, выровненные в середине заданного интервала. Один из простейших методов механического сглаживания — это метод простой скользящей средней. Для временного ряда $y_1, y_2, y_3, \dots, y_n$ определяется интервал сглаживания $m < n$. Среднее арифметическое для первых m значений вычисляется по формуле:

$$\bar{X}_t = \frac{1}{m} \sum_{k=t-p}^{t+p} x_k, \quad t > p,$$

где $\bar{X}_t$ — сглаженное значение, m — длина интервала сглаживания, а p — количество уровней, учитываемых при вычислении среднего. Ограничения этого метода связаны с тем, что первые и последние p уровней не могут быть сглажены, и он подходит только для линейных трендов.

Для более сложных трендов используется метод взвешенной скользящей средней, в котором каждому уровню присваиваются разные веса. Формула для вычисления взвешенной арифметической средней выглядит следующим образом:

$$\bar{X}_t = \frac{\sum_{k=t-p}^{t+p} p_k x_k}{\sum_{k=t-p}^{t+p} p_k},$$

где p_k — веса, определяемые с помощью метода наименьших квадратов.

Метод экспоненциального сглаживания выделяется тем, что в процессе вычисления сглаженного уровня используются только предшествующие

значения временного ряда, при этом каждому значению присваивается вес, уменьшающийся с увеличением временного расстояния от текущего момента.

Процесс экспоненциального сглаживания описывается рекуррентным соотношением:

$$S_t = \alpha y_t + (1 - \alpha)S_{t-1},$$

где α — параметр сглаживания, принимающий значения в интервале $0 < \alpha < 1$. Величина $\beta = 1 - \alpha$ обозначается как коэффициент дисконтирования.

Сглаженное значение можно представить как:

$$S_t = \alpha \sum_{i=0}^{t-1} (1 - \alpha)^i y_{t-i} + (1 - \alpha)^t S_0,$$

где S_0 — начальные условия.

Оптимальный выбор параметра α критически влияет на точность прогнозов. Варианты для оценки α включают:

$$\alpha = \frac{2}{n + 1},$$

и

$$\alpha \approx \frac{\sigma_n}{\sigma_\varepsilon},$$

где σ_n и σ_ε — среднеквадратические ошибки модели и исходного ряда соответственно .

Начальное значение S_0 устанавливается как:

$$S_0 = \frac{y_1 + y_2 + y_3}{3},$$

Метод экспоненциального сглаживания позволяет использовать весь временной ряд, сохраняя начальные и конечные значения, что делает его эффективным для прогнозирования спроса в розничной торговле.

Модель ARMA является одним из наиболее распространённых методов анализа временных рядов. Она используется для описания слабо стационарных стохастических процессов, объединяя авторегрессию (AR) и скользящее среднее (MA). AR-компонента предполагает, что текущее значение ряда за-

висит от своих прошлых значений, тогда как МА-компонента моделирует текущую ошибку как сумму прошлых шумовых вкладов.

Модель ARMA порядка (p, q) задаётся формулой:

$$X_t = \varepsilon_t + \sum_{i=1}^p \varphi_i X_{t-i} + \sum_{i=1}^q \theta_i \varepsilon_{t-i}.$$

Здесь X_t — значение ряда в момент времени t , ε_t — случайная ошибка, φ_i и θ_i — коэффициенты авторегрессии и скользящего среднего соответственно.

Построение модели ARMA включает три основных этапа:

1. **Идентификация:** выбор значений p и q для модели ARMA, что может потребовать корректировки.
2. **Оценивание:** уточнение параметров модели с использованием статистических методов и проверка гипотез.
3. **Диагностика:** анализ адекватности модели с использованием различных процедур проверки.

Модель ARMA позволяет учитывать влияние предыдущих значений временного ряда и исторических ошибок, объединяя преимущества обеих составляющих для более точного моделирования и прогнозирования.

Модель ARIMA разработана в рамках метода Бокса и Дженкинса и обеспечивает точные прогнозы, учитывая случайные колебания исторических данных. Модель $ARIMA(p, d, q)$ для нестационарного временного ряда X_t имеет вид:

$$\Delta^d X_t = c + \sum_{i=1}^p a_i \Delta^d X_{t-i} + \sum_{j=1}^q b_j \varepsilon_{t-j} + \varepsilon_t,$$

где ε_t — стационарный временной ряд, а c, a_i, b_j — параметры модели.

Параметры p, d и q обозначают:

- p — авторегрессионный компонент,
- d — уровень интегрированности,
- q — компонент скользящей средней.

При $d = 0$ модель становится обычной ARMA-моделью. Примером

ARIMA-модели является случайное блуждание:

$$x_t = x_{t-1} + \varepsilon_t \Rightarrow \Delta x_t = (1 - L)x_t = \varepsilon_t,$$

что представляет модель $ARIMA(0, 1, 0)$.

Временной ряд X_t называется интегрированным порядка k , если $X_t \sim I(k)$, когда разности порядка k являются стационарными. Порядок интегрированности соответствует d в модели $ARIMA(p, d, q)$.

Метод Бокс-Дженкинса включает три этапа:

1. Идентификация для определения значений p, d, q .
2. Оценка параметров через метод максимального правдоподобия.
3. Проверка диагностики остатков для адекватности модели.

Автокорреляция считается значительной, если превышает $\frac{2}{\sqrt{N}}$.

Модель SARIMA (Seasonal Autoregressive Integrated Moving Average) является универсальным инструментом прогнозирования временных рядов с выраженными сезонными характеристиками и представляет собой расширение модели ARIMA. Она включает авторегрессию (AR), интеграцию (I) и скользящее среднее (MA), а также сезонные компоненты для более точного моделирования сезонных колебаний.

Основные составляющие модели SARIMA:

- **Сезонный компонент (S)**: отражает периодичность, повторяющуюся с определённой частотой.
- **Авторегрессия (AR)**: моделирует связь текущего значения с прошлыми значениями.
- **Интеграция (I)**: дифференцирование данных для устранения нестационарных составляющих.
- **Скользящее среднее (MA)**: моделирует зависимость текущего значения от ошибок прошлых прогнозов.

Математически модель SARIMA описывается уравнением:

$$\Phi_p(B) \Phi_P(B^s) (1 - B)^d (1 - B^s)^D y_t = \Theta_q(B) \Theta_Q(B^s) \epsilon_t,$$

где y_t — наблюдаемое значение, B — оператор запаздывания, $\Phi_p(B)$ и $\Theta_q(B)$ — полиномы AR и MA соответственно, а параметры p, d, q, P, D, Q отвечают

за краткосрочные и сезонные компоненты.

Таким образом, SARIMA обеспечивает гибкость и эффективность в прогнозировании сложных сезонных временных рядов.

Четвертый раздел «Статистические методы в прогнозировании спроса» включает два раздела: «Оценка и точность модели» и «Оценка точности прогноза».

Оценка и точность модели включает в себя анализ графиков ACF и PACF и метод максимального правдоподобия.

Автокорреляция ($\rho(k)$) измеряет сходство между временным рядом и его запаздыванием, и рассчитывается как:

$$\rho(k) = \frac{\text{Cov}(X_t, X_{t-k})}{\sigma(X_t) \cdot \sigma(X_{t-k})},$$

где Cov — ковариация, σ — стандартное отклонение, X_t — значение в момент времени t .

Графики автокорреляционной функции (АКФ) помогают в выборе параметров модели, указывая на возможные значения p и q . Частичная автокорреляция ($\phi_{k,k}$) измеряется как:

$$\phi_{k,k} = \text{corr}(z_{t+k} - \hat{z}_{t+k}, z_t - \hat{z}_t),$$

где $\hat{z}$ — линейные комбинации предыдущих значений.

Полезность ACF и PACF в прогнозировании спроса:

- ACF выявляет сезонные колебания и паттерны, PACF уточняет влияние запаздываний.
- Позволяют определить порядок модели ARIMA, выбирая параметры p и q .
- Оценивают стационарность временного ряда, что необходимо для ARIMA.
- Улучшают точность прогнозов спроса в розничной торговле.
- Позволяют быстро адаптироваться к изменениям на рынке.

Метод максимального правдоподобия (ММП) используется для оценки параметров модели ARIMA (p, d, q) путём максимизации функции правдоподобия:

$$\ln L(a, \beta, \sigma^2) = -\frac{1}{2}n \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=1}^n (X_t - \hat{X}_t)^2,$$

где X_t — фактические значения, $\hat{X}_t$ — предсказанные моделью, a_i и β_j — коэффициенты, σ^2 — дисперсия ошибок, n — количество наблюдений.

Для нахождения максимума функции правдоподобия вычисляются её частные производные по параметрам a_i, β_j, σ^2 . Процесс продолжается до тех пор, пока изменения логарифмической функции не станут незначительными (например, меньше 10^{-4}). Оцененные параметры записываются как:

$$(\hat{a}_i, \hat{\beta}_j, \hat{\sigma}^2) = \arg \max a_i, \beta_j, \sigma^2 \in Q \ln L(a_i, \beta_j, \sigma^2).$$

ММП позволяет точно определить порядок авторегрессии (p), интеграции (d) и скользящего среднего (q), что критично для успешного прогнозирования спроса в розничной торговле

Для оценки точности прогнозов модели ARIMA используются различные метрики:

1. Информационный критерий Акаике (AIC).

$$AIC = -2l + 2k,$$

где l — логарифм функции правдоподобия, k — количество параметров. Меньшие значения AIC указывают на лучшую модель.

2. Среднеквадратичная ошибка (MSE).

$$MSE = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{X}_i)^2,$$

где X_i — реальные значения, $\hat{X}_i$ — предсказанные. Низкие значения MSE означают высокую точность.

3. Средняя абсолютная процентная ошибка (MAPE).

$$MAPE = 100 \cdot \frac{1}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right|,$$

где A_t — фактические, F_t — предсказанные значения. Метрическая интерпретация точности в процентах.

4. Коэффициент детерминации (R^2).

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{X}_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2},$$

где $\bar{X}$ — среднее значение зависимой переменной. Значение ближе к 1 указывает на хорошую объясняющую способность модели.

5. Корень средней квадратичной ошибки (RMSE).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \hat{X}_i)^2},$$

где X_i — наблюдаемые, $\hat{X}_i$ — предсказанные значения. Низкие значения RMSE свидетельствуют о большей точности.

Эти метрики помогают оценить эффективность модели ARIMA в прогнозировании спроса, учитывая сезонные и трендовые паттерны, что особенно актуально в условиях изменчивого рынка.

Практическая часть включает в себя следующие разделы: «Модели для стационарного ряда» и «Модели для нестационарного ряда».

Для анализа и прогнозирования спроса в розничной торговле в рамках стационарного ряда используется датасет «Retail Store Inventory Forecasting» с платформы Kaggle, содержащий более 73 000 строк ежедневных записей по продажам, запасам и внешним факторам.

Для анализа будет агрегирована выручка от продаж электроники в магазине S001, рассчитываемая как произведение количества проданных единиц на цену. Проведён анализ статистических характеристик выручки, включая проверку на наличие пропущенных значений (не обнаружены) и обработку выбросов (заменены медианным значением).

Для проверки стационарности временного ряда применён тест Дики-Фуллера. Далее построены функции ACF и PACF для исходного временного ряда без выбросов. После анализа показателей и функций ACF и PACF был построен прогноз с использованием модели ARIMA(1, 0, 1). Статистиче-

ская сводка модели содержит 574 наблюдения, логарифм правдоподобия = -3569.257, AIC = 7146.515 и BIC = 7163.925. Константа значима ($p < 0.000$), однако коэффициенты AR(1) и MA(1) нестатистически значимы ($p = 0.119$ и 0.261).

Визуализация результатов прогноза была осуществлена с помощью графика, который представлен на рисунке 1.

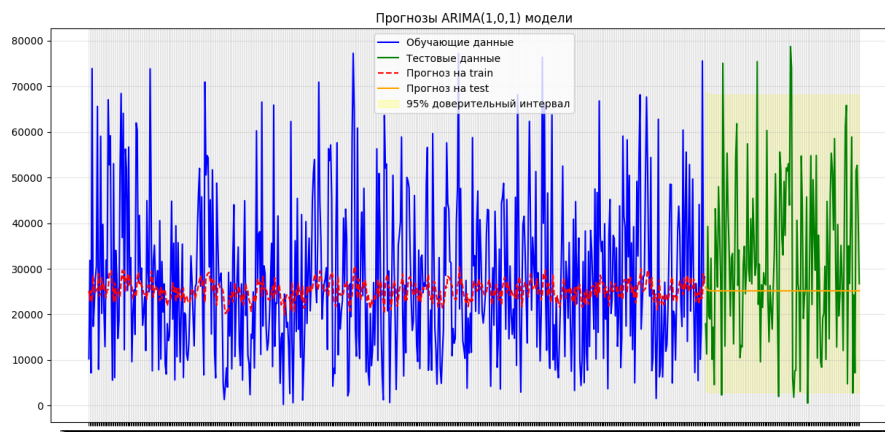


Рисунок 1 – Прогноз ARIMA (1,0,1)

В разделе «Модели для нестационарного ряда» будут использованы данные о розничной торговле в США, полученные с сайта Федерального резервного банка Сент-Луиса. Рассматриваются предварительные данные о розничных продажах, обозначенные как розничная торговля (RSXFSN). Данные включают в себя продажи в сфере электронной коммерции и представляют ежемесячные значения, что позволяет анализировать тренды и сезонные колебания.

Набор данных охватывает период с января 1992 года по настоящее время, что позволяет увидеть изменения в объеме розничных продаж более чем за 30 лет.

Для нестационарного ряда были построены различные модели ARIMA и SARIMA, включая модели с автоматическим подбором параметров. Каждая модель была проанализирована с помощью статистической сводки, которая включала такие параметры, как AIC, BIC и логарифм правдоподобия.

После этого было проведено сравнение моделей по качеству прогнозирования, что позволило выявить наиболее эффективную модель. Прогнозы для каждой модели были построены и проанализированы, что дало возможность

оценить их точность и способность к прогнозированию будущих значений розничных продаж.

Прогноз ARIMA (1,1,1) для нестационарного ряда на обучающей и тестовой выборке представлен на рисунке 2.

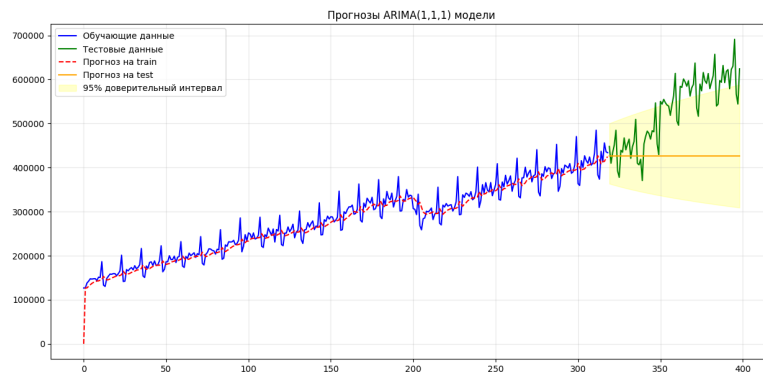


Рисунок 2 – Прогноз ARIMA (1,1,1)

Прогноз SARIMA (1,1,1)(1,1,1,12) для нестационарного ряда на тестовой выборке представлен на рисунке 3.

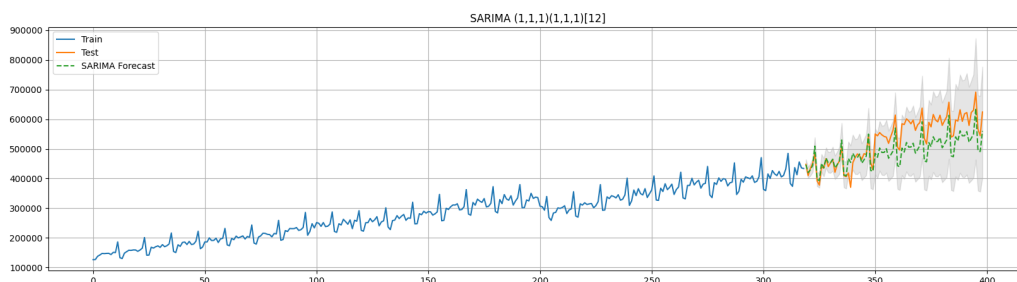


Рисунок 3 – Прогноза SARIMA на тестовой выборке (1,1,1)(1,1,1,12)

Прогноз SARIMA (0,1,1)(2,1,2,12) для нестационарного ряда на тестовой выборке представлен на рисунке 4.

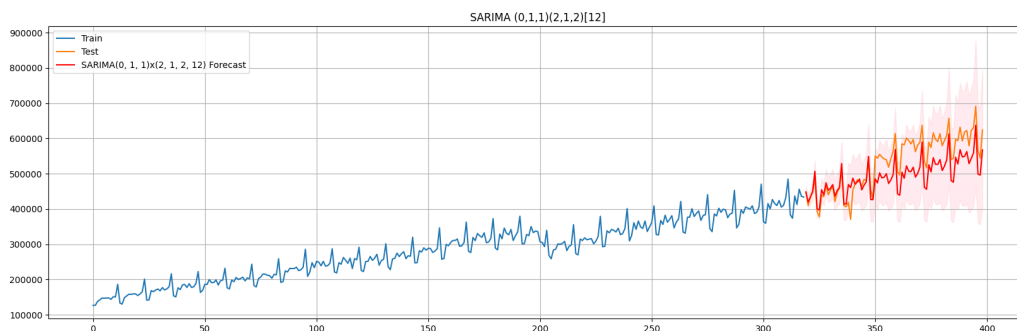


Рисунок 4 – Прогноз SARIMA (0,1,1)(2,1,2,12) на тестовой выборке

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

В ходе выполнения дипломной работы была достигнута цель — разработка и сравнение моделей прогнозирования спроса на основе временных рядов для розничной торговли, а также создание универсального шаблонного кода для анализа различных временных рядов. Построены модели ARIMA для стационарного ряда и SARIMA для нестационарных, учитывающих сезонность и тренд. Проведен анализ качества прогноза с использованием статистических метрик и визуализации, что показало преимущества SARIMA в учёте сезонных эффектов.

Ключевые этапы работы включали:

1. Исследование временных рядов и обоснование выбора моделей.
2. Построение и валидация моделей ARIMA и SARIMA с использованием диагностических процедур.
3. Визуализация результатов для иллюстрации прогнозов.
4. Разработка универсального шаблонного кода, применимого к различным временным рядам.

Результаты работы могут служить основой для дальнейших исследований и практических применений в анализе и прогнозировании временных рядов.