Санкт-Петербургский государственный университет

Математическое обеспечение и администрирование информационных систем
Информационно-аналитические системы

Григорович Роман Владимирович

Предсказание количества безработных в регионе: разработка критериев для модели.

Бакалаврская работа

Научный руководитель:
к. ф.-м. н., доцент Графеева Н. Г.

Рецензент:
Калугин Д. И.

Санкт-Петербург
2017

SAINT-PETERSBURG STATE UNIVERSITY

Software and Administration of Information Systems
Analytical Information Systems

Roman Grigorovich

Prediction of the number of unemployed in the region: development of criteria for the model

Bachelor's Thesis

Scientific supervisor:
associate professor Natalia Grafeeva

Reviewer:
Dmitry Kalugin

Saint-Petersburg
2017

Оглавление

Введение **5**

1. Постановка задачи **6**

2. Существующие исследования **7**

2.1. Описание моделей 7

2.1.1. Модель Хольта 7

2.1.2. Модель Хольта-Уинтерса 8

2.1.3. Линейная регрессия 8

2.1.4. Модель авторегрессии - скользящего среднего . . 9

2.1.5. Модель авторегресии и распределённого лага . . . 10

2.1.6. Нейронные сети прямого распространения 11

2.1.7. Рекуррентные нейронные сети 11

3. Сравнение моделей **12**

4. Исходные данные **13**

5. Построение моделей без предикторов **14**

5.1. Модель Хольта 14

5.2. Модель Хольта-Уинтерса 15

5.3. Модель авторегрессии - скользящего среднего 16

6. Построение моделей с предикторами **17**

6.1. Линейная регрессия 18

6.2. Модель авторегрессии и распределённого лага 19

6.3. Модели, использующие нейронные сети 20

6.3.1. Нейронные сети прямого распространения 20

6.3.2. Рекуррентная нейронная сеть 21

7. Результаты **22**

Заключение **23**

Список литературы **24**

Введение

Одной из важнейших проблем современной экономики является безработица. Она представляет собой сложное и противоречивое макроэкономическое явление экономической жизни.

Безработица — наличие в стране людей, составляющих часть экономически активного населения, которые способны и желают трудиться по найму, но не могут найти работу.

На уровень безработицы оказывают влияние различные факторы, такие как: уровень заработной платы, уровень налогов, страхование на случай безработицы, ВВП. Например, известен закон Оукена [1], который говорит о том, что снижение темпа роста ВВП на 2 % приводит к повышению уровня безработицы на 1 %. К последствиям безработицы относятся:

- Снижение доходов;
- Потеря квалификации;
- Экономические последствия;
- Ухудшение криминогенной ситуации;
- Ухудшение динамики роста интереса населения к труду;
- Снижение уровня обеспеченности домохозяйств.

1. Постановка задачи

Целью данной работы является предсказание количества безработных на основе исторических данных о числе безработных, а также - других экономичеких показателей.

Для достижения цели были поставлены следующие задачи:

- Изучение предметной области;
- Сбор данных для исследования;
- Применение различных моделей предсказания;
- Сравнение моделей друг с другом.

2. Существующие исследования

Модели прогнозирования можно разделить на две группы: модели, основывающиеся только на исторических данных; модели, использующие кроме исторических данных дополнительные факторы.

К первой группе относятся модели, основанные на сглаживании временного ряда, такие как: экспоненциальное сглаживание, метод Хольта, метод Хольта-Уинтерса. [3] [9] [11] А также авторегрессионные модели: модель авторегрессии-скользящего среднего (ARMA), авторегрессионная условная гетероскедастичность (ARCH, GARCH). [8] [2] [5]

Ко второй группе - модели, использующие нейронные сети (сети прямого распространения, рекуррентные нейронные сети с длительной краткосрочной памятью, сети радиально-базисных функций); множественная регрессия, модель авторегрессии и распределённого лага. [13] [7] [12]

2.1. Описание моделей

2.1.1. Модель Хольта

Существуют наивные методы прогнозирования такие как: скользящая средняя, взвешенная средняя, простое экспоненциальное сглаживание. Эти методы не позволяют сделать прогноз долгосрочным - для получения прогноза мы должны знать фактическую величину предыдущего значения. Поэтому рассмотрим расширение этих методов. Ряд разбивается на составляющие - уровень l и тренд b.

$l_t = \alpha y_t + (1 - \alpha)(l_{t-1} + b_{t-1})$

$b_t = \beta(l_t - l_{t-1}) + (1 - \beta)b_{t-1},$

$\hat{y}_{t+1} = l_t + b_t$

Параметр α характеризует сглаживание ряда вокруг тренда, параметр β — сглаживание тренда. Чем меньше значения, тем меньший вес будет приписываться последним наблюдениям и тем более сглаженным окажется построенный ряд.

2.1.2. Модель Хольта-Уинтерса

Идеей метода является добавление ещё одной составляющей составляющей — сезонности s.

$l_t = \alpha(y_t - s_{t-L}) + (1-\alpha)(l_{t-1} + b_{t-1})$

$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$

$s_t = \gamma(y_t - l_t) + (1-\gamma)s_{t-L}$

$\hat{y}_{t+m} = l_t + mb_t + s_{t-l+1+(m-1)modL}$

Уровень l определяется текущим значением ряда с вычтенной сезонной составляющей. Сезонная часть определяется текущим значением ряда с вычтенным уровнем и предыдущим значением составляющей.

2.1.3. Линейная регрессия

Модель вида $y = f(x, b) + \varepsilon$, где b - параметры модели, ε - случайная ошибка модели, называется линейной регрессией, если функция регрессии $f(x, b)$ имеет вид: $f(x, b) = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k$, где b_j - коэффициенты регрессии, x_j - факторы модели, k - количество факторов.

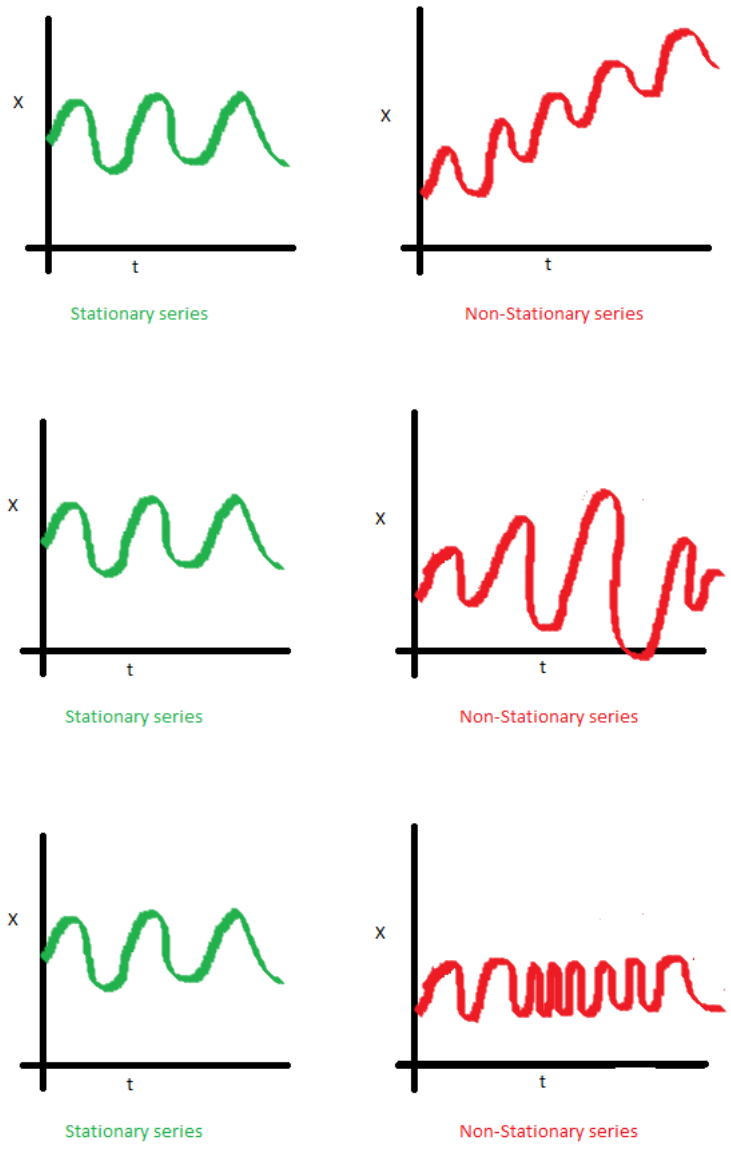
Если количество факторов больше одного, то такую модель называют множественной регрессией.

2.1.4. Модель авторегрессии - скользящего среднего

Данную модель можно применять, если прогнозируемый ряд является стационарным.

Под стационарностью понимают свойство процесса не менять своих статистических характеристик с течением времени.

Примеры стационарных и нестационарных рядов:



На рисунке 1 временной ряд справа нестационарный, так как у него увеличивается математическое ожидание.

На рисунке 2 - у ряда справа непостоянная дисперсия (в разные периоды изменяется разброс значений ряда).

На рисунке 3 - непостоянная ковариация (значения ряда местами сближаются).

Стационарность важна, так как по ряду с таким свойством проще строить прогноз. Статистические показатели ряда не будут изменяться со временем.

Существует несколько тестов на стационарность. Один из них - расширенный тест Дикки-Фуллера [6].

Модель авторегрессии - скользящего среднего (ARMA) - моделя для прогнозирования временных рядов. Данная модель обобщает две другие модели - модель авторегрессии (AR) и модель скозьзящего среднего (MA).

Пусть задан временной ряд: $X_1, X_2, \ldots, X_i$. Нужно построить прогноз.

$y_t = c + \varepsilon_t + \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{i=2}^{q} \beta_i \varepsilon_{t-i}$,

c - константа,

ε_t - белый шум,

$\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q$ - действительные числа, коэффициенты авторегрессии и коэффициенты скользящего среднего, соответственно.

Данная модель может рассматриваться как модель регрессии, в которой в качестве факторов используются прошлые значения этого ряда.

2.1.5. Модель авторегресии и распределённого лага

ADL (autoregressive distributed lags) модель - модель, в которой текущие значения ряда зависят от прошлых значений этого ряда и от текущих и прошлых значений других рядов.

Модель ADL(p,q) с одной экзогенной переменной имеет вид:

$y_t = a_0 + \sum_{i=1}^{p} a_i y_{t-i} + \sum_{j=0}^{q} b_j x_{t-j} + \varepsilon_t$,

p - количество лагов зависимой переменной,

q - количество лагов предиктора.

Модель можно обобщить на несколько экзогенных переменных.

2.1.6. Нейронные сети прямого распространения

С помощью нейронных сетей возможно прогнозирование временных рядов, поскольку у нейронной сети есть способноть обобщать и выделять скрытые зависимости между данными, поступающими на вход и выходными данными. Сеть по окончанию обучения может предсказывать значения ряда, основываясь на предыдущих значениях и значениях факторов в настоящий момент времени.

Многослойный перцептрон является сетью прямого распространения. Сигналы в такой сети передаются в одном направлении от входных слоёв к выходным. Основными компонентами многослойного перцептрона являются: входные узлы, образующие входной слой, один или несколько скрытых слоёв, один выходной слой. В работах [12] [10] использовались архитектуры с одним скрытым слоем.

2.1.7. Рекуррентные нейронные сети

В отличие от сетей прямого распространиния, рекуррентные нейронные сети имеют обратные связи. В таких сетях нейроны обмениваются информацией между собой. Нейрон получает помимо входных данных информацию о предыдущем состоянии сети. Тем самым у сети появляется память, что позволяет анализировать данные, в которых важен порядок значений, такие как временные ряды.

3. Сравнение моделей

Помимо построения самих моделей, одной из задач данной работы является выявление лучшей модели, которая способна показать меньшую ошибку. Одни из наиболее частых показателей, которые используются для сравнения моделей различных типов - средняя процентная ошибка модели (MAPE) и средняя квадратичная ошибка модели (MSE).

Оценка MAPE применяется для временных рядов, фактические значения которых значительно больше 1

$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{y}_i - y|}{y_i}$

$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$

Коэффициент детерминации (R^2) - это доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью.

$R^2 = 1 - \frac{SSE}{TSS}$,

где $SSE = \sum_{i=1}^{N}(\hat{y}_i - y_i)^2$ - это сумма квадратов ошибок модели, $TSS = \sum_{i=1}^{N}(y_i - \bar{y_i})^2$ - сумма квадратов отклонений фактических значений от средней величины.

Для того, чтобы можно было сравнивать модели с разным числом факторов так, чтобы число факторов не влияло на R^2 обычно используется скорректированный коэффициент детерминации:

$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-k}$,

где n - число наблюдений, а k - число факторов. У коэффициента детерминации есть важный недостаток. Он расчитывается по обучающей выборке, а значит показывает только лишь то насколько хорошо описываются данные.

4. Исходные данные

В качестве данных используются временные ряды из открытой базы Federal Reserve Economic Data [4]. Взяты ежемесячеые данные с 1949 по 2016г.

- Количество безработных;
- Денежная база;
- Средняя заработная плата;
- Индекс инфляции;
- Валовой внутренний продукт;
- Пособие по безработице;
- Цена корпоративных облигаций;
- Цена на нефть.

Исходный ряд с количеством безработных разбит на обучающую и тестовую выборки.

Обучающая выборка с 1949 по 2014 год. Тестовая выборка с 2015 по 2016 год.

5. Построение моделей без предикторов

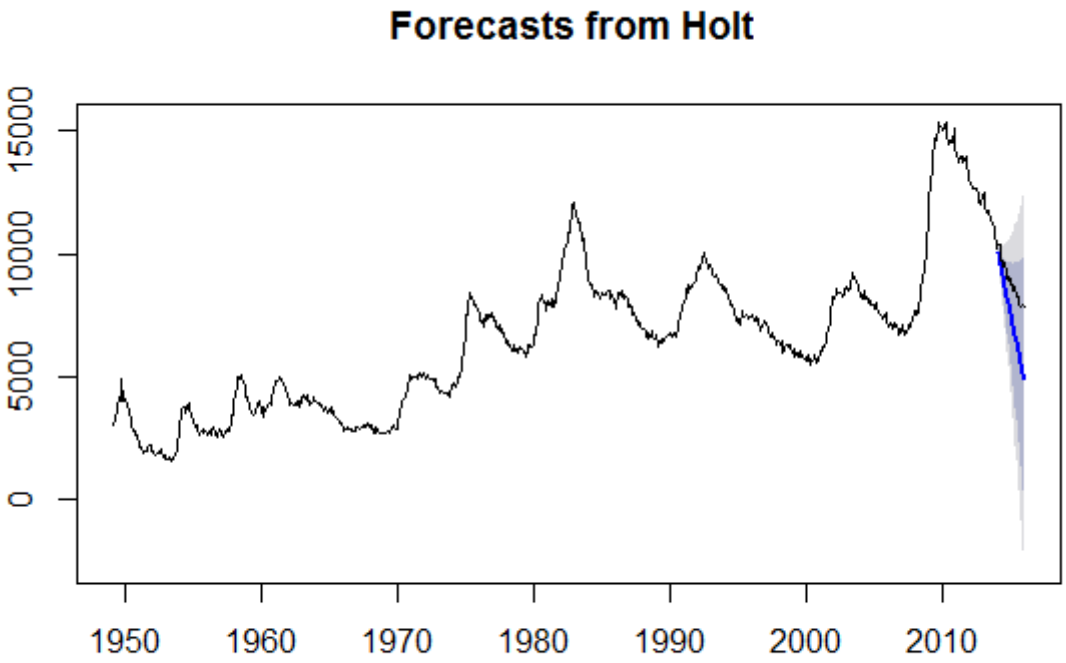
В данном разделе строятся модели прогнозирования, использующие один временной ряд, содержащий данные о количестве безработных.

5.1. Модель Хольта

Параметры модели подбирались так, чтобы минимизировать сумму квадратов ошибки и средний процент ошибки модели.

$\alpha = 0.31, \beta = 0.15$

$MSE = 2530998, MAPE = 0.164, R^2 = 0.56$

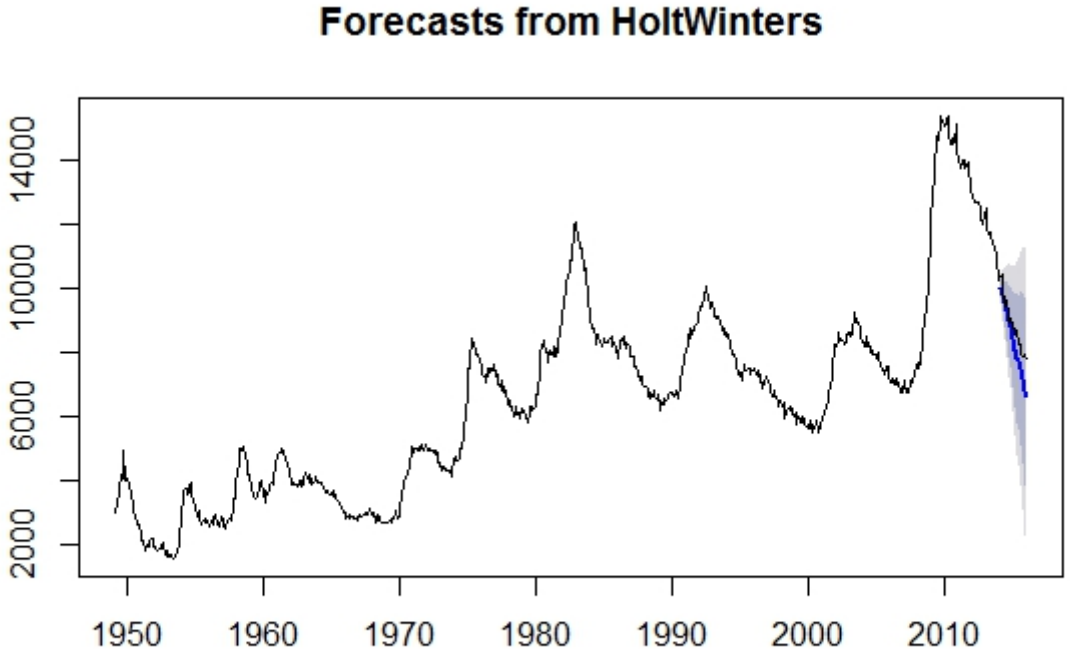


5.2. Модель Хольта-Уинтерса

Параметры модели подбирались так, чтобы минимизировать сумму квадратов ошибки и средний процент ошибки модели.

$\alpha = 0.215, \beta = 0.102, \gamma = 0.456$

$MSE = 321272, MAPE = 0.085, R^2 = 0.63$



5.3. Модель авторегрессии - скользящего среднего

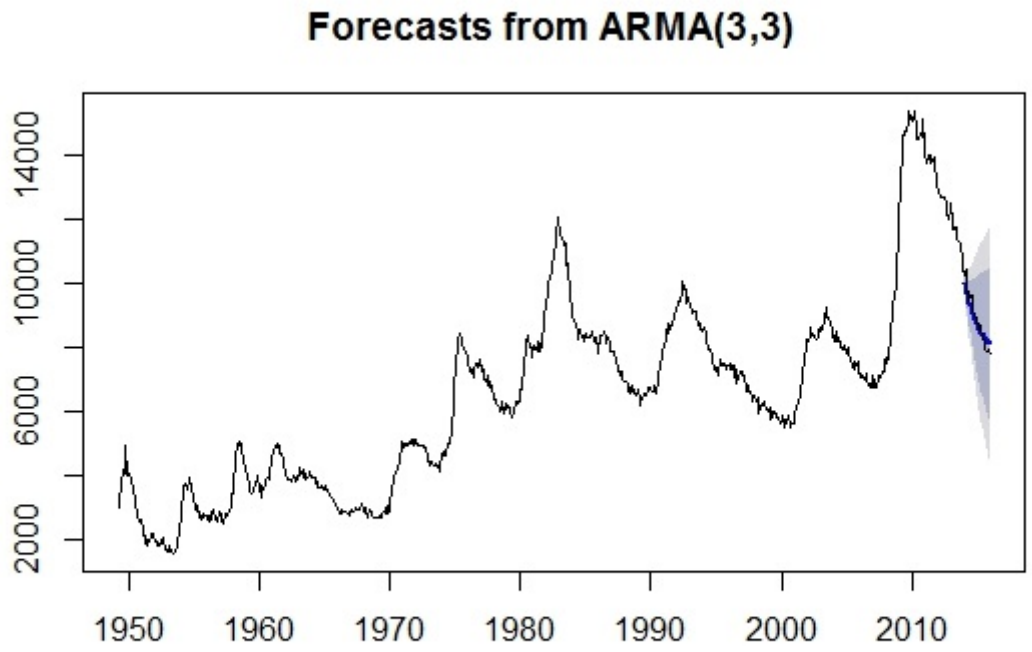
Исследуемый ряд удовлетворяет расширенному тесту Дикки-Фуллера, следовательно является стационарным.

Для построения ARMA(p,q) модели необходимо определить подкласс, к которому будет относиться модель. В данном случае выбор подкласса значит выбор параметров p, q модели.

Коэффициенты p и q подбирались так, чтобы среднеквадратическая ошибка модели и средний процент ошибки модели были минимальны.

Наилучшей моделью оказалась ARMA(3,3). То есть для построения прогноза используются данные за последние 3 месяца.

$MSE = 295577, MAPE = 0.043, R^2 = 0.82$



6. Построение моделей с предикторами

В рассмотренных выше моделях использовались только значения одного временного ряда в прошлом.

При рассмотрении различных показателей в экономике нас интересует зависимость различных экономических величин. То есть, как текущее значение того или иного экономического показателя зависит от других показателей, а не только от его предыдущих значений.

В качестве таких показателей дальше будут рассматриваться:

- Денежная база;
- Средняя заработная плата;
- Индекс инфляции;
- Валовой внутренний продукт;
- Пособие по безработице;
- Цена корпоративных облигаций;
- Цена на нефть.

6.1. Линейная регрессия

Множественная регрессия позволяет исследовать влияние нескольких независимых переменных (предикторов) на одну зависимую переменную (количество безработных).

Сначала включим в модель все наши переменные (таблица 1).

Таблица 1: Модель со всеми предикторами

var	coefficients	p-value
Облигации	112.9968	0.0037969
Денежная база	-1.94468	0.000000001138876
Зарплата	1022.09318	0.0004524
ВВП	-1.4429	0.000000000000000222
Инфляция	8.4452734464	0.4483352
Пособие по безработице	0.00455	0.000000000000000222
Цены на нефть	34.89379	0.000000000094079
Исправленный R^2 : 0.491		

Индекс инфляции не является значимым. Исключим его из модели (Таблица 2).

Таблица 2: Модель без индекса инфляции

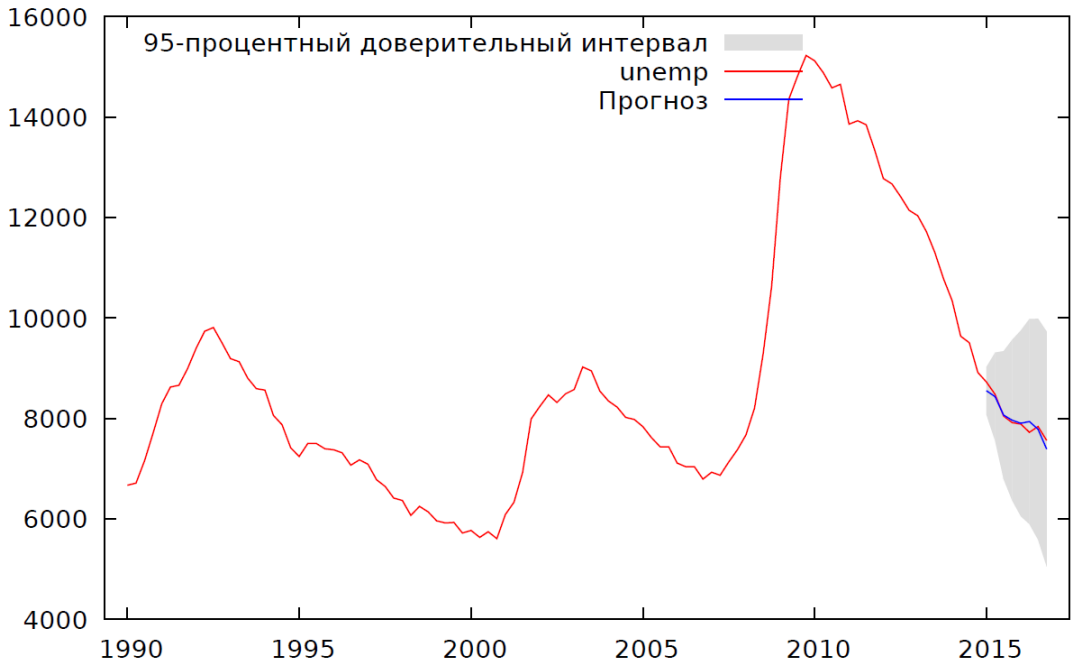
var	coefficients	p-value
Облигации	122.320	0.00096992
Денежная база	-2.0722	0.000000000000030944
Зарплата	1120.820	0.000017845537038594
ВВП	-1.3679	0.000000000000000222
Пособие по безработице	0.0046	0.000000000000000222
Цены на нефть	33.7038	0.000000000052805460
Исправленный R^2 : 0.523		

Теперь все предикторы являются значимыми. Таким образом, модель линейной регрессии без индекса инфляции является оптимальной моделью.

6.2. Модель авторегрессии и распределённого лага

Среди рассмотренных моделей с разными параметрами, лучшей оказалась модель ADL(3,2). То есть для построения прогноза используются данные о числе безработных за последние 2 месяца, а также данные обо всех экзогенных переменных за последние 3 месяца. Как и в ранее рассмотренных моделях минимизировалась среднеквадратическая ошибка.

$MSE = 290312, MAPE = 0.063, R^2 = 0.81$

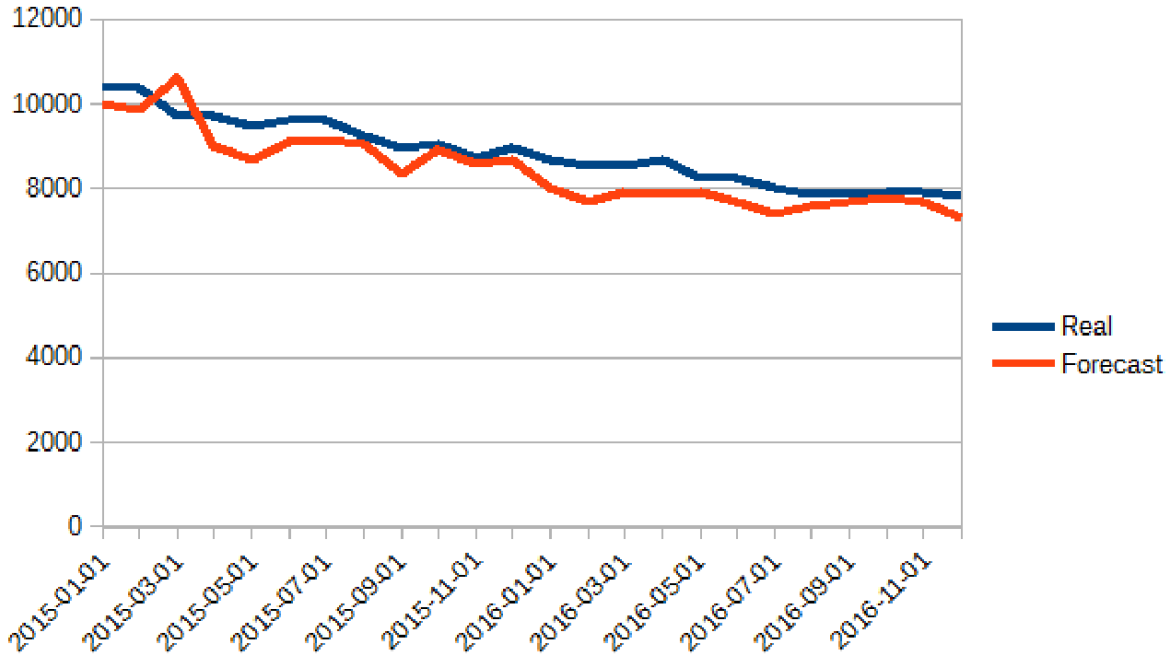


6.3. Модели, использующие нейронные сети

6.3.1. Нейронные сети прямого распространения

При построении модели были опробованы различные архитектуры сети. Наилучшей оказалась модель со следующей архитектурой: 7 нейронов на входном слое, один скрытый слой с 8 нейронами, один нейрон на выходе.

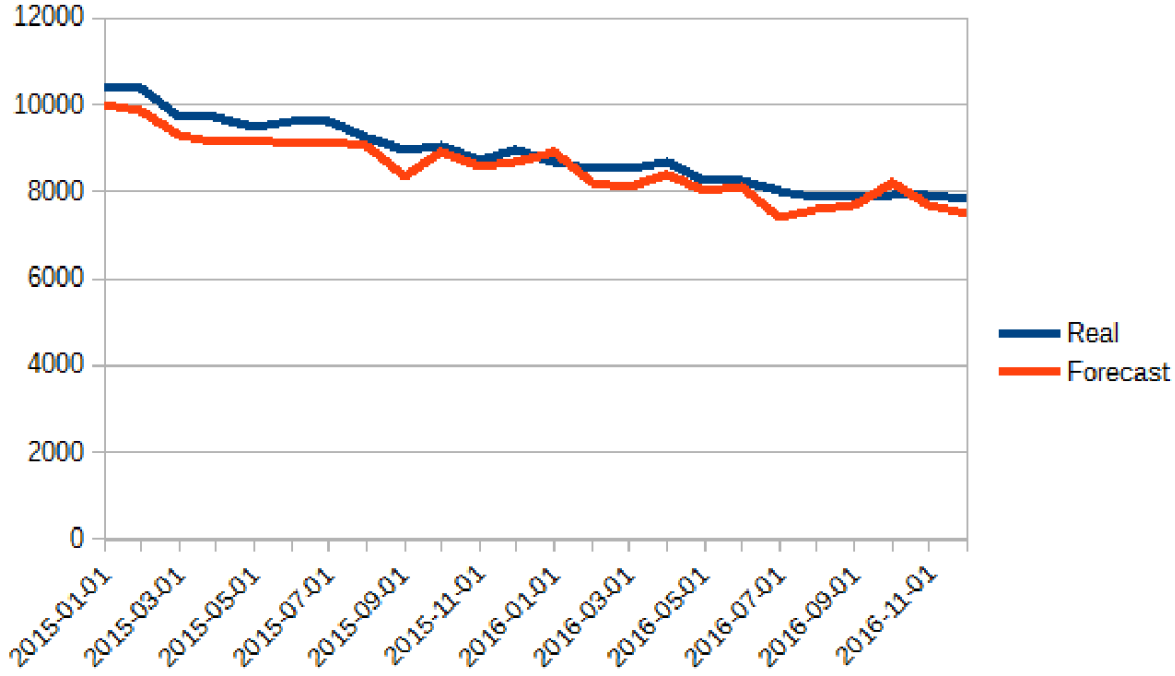
$MSE = 281899.1917, MAPE = 0.055, R^2 = 0.80$



6.3.2. Рекуррентная нейронная сеть

При построении модели были опробованы различные архитектуры сети. Наилучшей оказалась модель со следующей архитектурой: 7 нейронов на входном слое, один скрытый слой с 6 нейронами, один нейрон на выходе.

$MSE = 135539.5027, MAPE = 0.03, R^2 = 0.92$



7. Результаты

В таблице представлены оценки качества построенных моделей.

Model	Holt	Holt-Winters	LR	ARMA	ADL	MLP	RNN
MSE	2530998	321272	341289	295577	290312	281899	135539
MAPE	0.164	0.085	0.18	0.043	0.063	0.055	0.03
R^2	0.56	0.63	0.523	0.81	0.84	0.80	0.92

Исходя из таблицы, наилучшей моделью, построенной в данной работе оказась модель на основе рекуррентных нейронных сетей. Эта модель дала наименьшую абсолютную процентную ошибку 3% и она объясняет 92% дисперсии модели. Это объясняется её способностью запоминать данные на каждой итерации.

Также достаточно точный прогноз дала модель ARMA. Процентная ошибка, которой составила 4%. Объясняет модель ARMA 82% дисперсии модели.

Заключение

В данной работе для прогнозирования количества безработных были рассмотрены классические модели прогнозирования временных рядов (модели ARIMA, ADL, Holt, Holt-Winters, линейная регрессия), а также методы, основанные на применении нейронных сетей: многослойный перцептрон и рекуррентная нейронная сеть.

Для оценки моделей строился прогноз на 2 года вперед. Были измерены средняя квадратичная ошибка модели и средний процент ошибки модели.

По результатам можно сказать, что модели, основанные на нейронных сетях справляются с прогнозированием временных рядов не хуже, чем классические модели прогнозирования.

Список литературы

[1] Ball Laurence, Jalles João Tovar, Loungani Prakash. Do Forecasters Believe in Okun's Law? An Assessment of Unemployment and Output Forecasts. — International Monetary Fund, 2014.

[2] Brailsford Timothy J., Faff Robert W. An evaluation of volatility forecasting techniques. — Journal of Banking & Finance, 1996.

[3] Dumičić Ksenija, Čeh Časni Anita, Žmuk Berislav. Forecasting Unemployment Rate in Selected European Countries Using Smoothing Methods. — International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering, 2015.

[4] Federal Reserve Economic Data. — URL: `https://fred.stlouisfed.org/`.

[5] Floros Christos. Forecasting the UK Unemployment rate: model comparisons. — International Journal of Applied Econometrics and Quantitative Studies. Vol.2, 2005.

[6] Greene William H. Econometric Analysis. — Macmillan Publishing Company, 1997.

[7] Hansen Bruce E. Time Series Econometrics for the 21st Century. — Updating the Undergraduate Econometrics Curriculum, 2016.

[8] KURITA Takamitsu. A Forecasting Model for Japan's Unemployment Rate. — Eurasian Journal of Business and Economics, 2010.

[9] Nasir Mohd Nadzri Mohd, Hwa Kon Mee, Mohammad Huzaifah. An Initial Study on the Forecast Model for Unemployment Rate. — The Daily, 2007.

[10] Sharma Saloni, Singh Sanjay. Unemployment Rates Forecasting Using Supervised Neural Networks. — Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference, 2016.

[11] Sugiarto Vicky Chrystian, Sarno Riyanarto, Sunaryono Dwi. Sales Forecasting Using Holt-Winters in Enterprise Resource Planning At Sales and Distribution Module. — International Conference on Information & Communication Technology and Systems, 2016.

[12] Wang Guangming, Zheng Xiangna. The Unemployment Rate Forecast Model Based on Neural Network. — International Workshop on Intelligent Systems and Applications, 2009.

[13] Zhang Guoqiang, Patuwo B. Eddy, Hu Michael Y. Forecasting with artificial neural networks:: The state of the art. — International Journal of Forecasting, 1998.