Отчет по результатам практической работы

1.Описание задачи

Описание:

В файле «data_for_test.xlsx» представлены временные ряды предикторов и целевого показателя (суммарное количество телефонных звонков возможных покупателей в дилерские центры, реализующие автомобиль определённой марки).

Задача:

Построить регрессионную модель предсказания зависимой величины с помощью предоставленных факторов.

Датасет:

6 столбцов, 762 строки

Признаки:

`Date` – дата

`TV` – оценка числа контактов (в тысячах) с целевой аудиторией через ТВ-рекламу

`OOH` – оценка числа контактов (в тысячах) с целевой аудиторией через наружную рекламу за указанный месяц

`Seasonal_Sales` – оценка сезонной составляющей продаж автомобилей изучаемой марки в помесячной детализации (от 0 до 1)

`Usd_rate` – курс доллара по Yahoo Finance (рублей за 1 доллар)

`Y` – количество звонков (целевая переменная)

Типы данных:

`Date` - datetime, `TV` - int, остальные – float

	Date	Y	TV	OOH	Seasonal_Sales	Usd_rate
0	2013-01-01	NaN	0	0.0	0.060802	30.502001
1	2013-01-02	0.0	0	0.0	0.060802	30.337200
2	2013-01-03	17.0	0	0.0	0.060802	30.156500
3	2013-01-04	17.0	0	0.0	0.060802	30.271000
4	2013-01-05	8.0	0	0.0	0.060802	30.271000

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 762 entries, 0 to 761
Data columns (total 6 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Date            762 non-null    datetime64[ns]
 1   Y               759 non-null    float64
 2   TV              762 non-null    int64
 3   OOH             762 non-null    float64
 4   Seasonal_Sales  762 non-null    float64
 5   Usd_rate        762 non-null    float64
dtypes: datetime64[ns](1), float64(4), int64(1)
memory usage: 35.8 KB
```

Есть 3 пропуска в целевой переменной. Обычное удаление пропусков можешь повлечь разрушение монотонности данных. Дальше посмотрим, можно ли их чем-то заменить.

Что можно сделать:

1) Сделать из `Data` индекс
2) Преобразовать тип данных в `Y` из float в int, т.к. по описанию данных это количество звонков - значение всегда должно быть целочисленным.

2. Результаты исследовательского анализа данных

Date – Дата

Даты последовательны и не повторяются.

Первая и последняя дата соответствуют заявленным в описании.

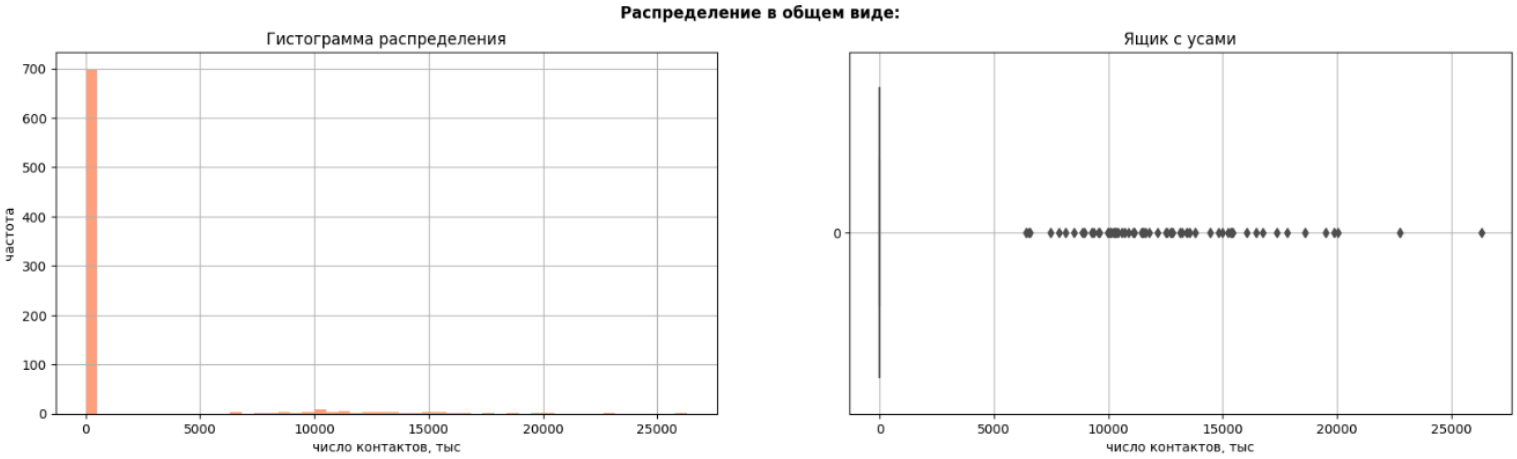
Даты идут подряд от 01.01.2013 до 01.02.2015 без перерыва.

```
print ('Первая дата:', df.index.min())
print ('Последняя дата:', df.index.max())
print ('Временной отрезок:', df.index.max() - df.index.min())
```

```
Первая дата: 2013-01-01 00:00:00
Последняя дата: 2015-02-01 00:00:00
Временной отрезок: 761 days 00:00:00
```

TV – показатель медиаактивности на ТВ

Всего в 8.27% датасета есть информация о ненулевой оценке числа контактов через ТВ, в остальном это значение равно нулю:

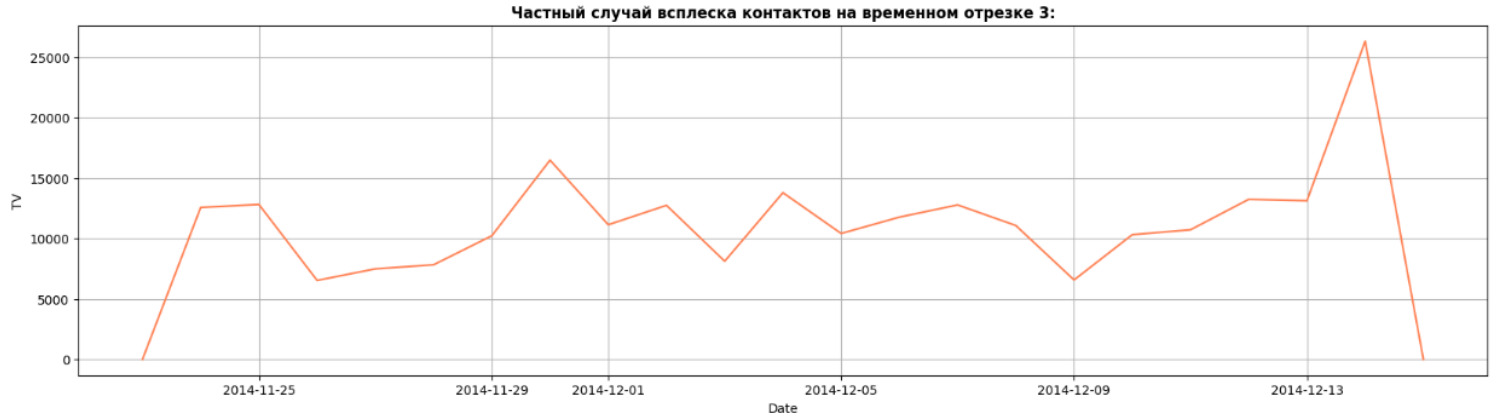


Наблюдается 3 всплеска активности контактов, пришедших через ТВ:

1) 2013: 17 февраля - 11 марта, - от 5000 до 22000 тыс контактов

2) 2014: 23 июня - 13 июля, - от 10000 до 20000 тыс контактов

3) 2014: 24 ноября - 14 декабря, - от 7000 до 26000 тыс контактов



- Наибольший разброс и наивысшее количество контактов в день видим в ноябре-декабре 2014 (последний график).
- Пик - 14 декабря 2014 (последний график).
- Наибольший минимум контактов в день - июнь-июль 2014 (2 график).

Изменение степени влияния рекламной активности на ТВ:

Корреляция TV и Y:

январь-август 2013: 0.284

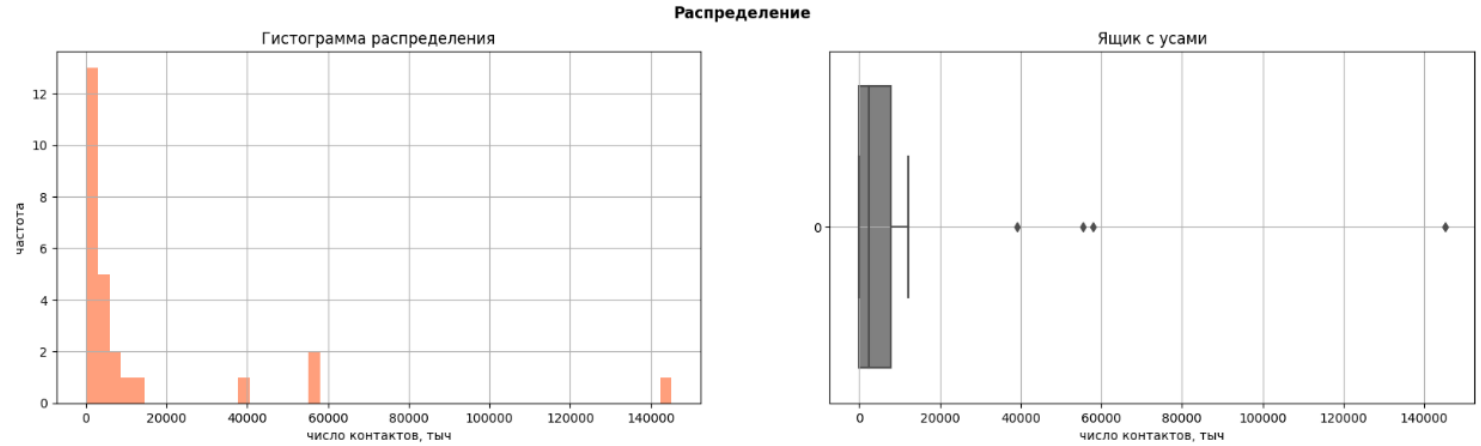
май-сентябрь 2014: 0.0

октябрь 2014 - февраль 2015: 0.52

Октябрь 2014 – февраль 2015: корреляция наивысшая.
Январь-август 2013: эффективность рекламы в два раза ниже.
Май-сентябрь 2014: наименьшее влияние ТВ-рекламы на количество звонков.

OOH – Наружная реклама

В 42% месяцев в представленном датасете число контактов по наружной рекламе равно нулю:



	count	mean	std	min	25%	50%	75%	max
OOH	26.0	13734.923077	31365.322343	0.0	0.0	2460.75	7817.25	145137.0

Пики контактов:

2013: март; 2014: март, апрель, июль

Самое большое значение:

145.137 млн контактов в месяц

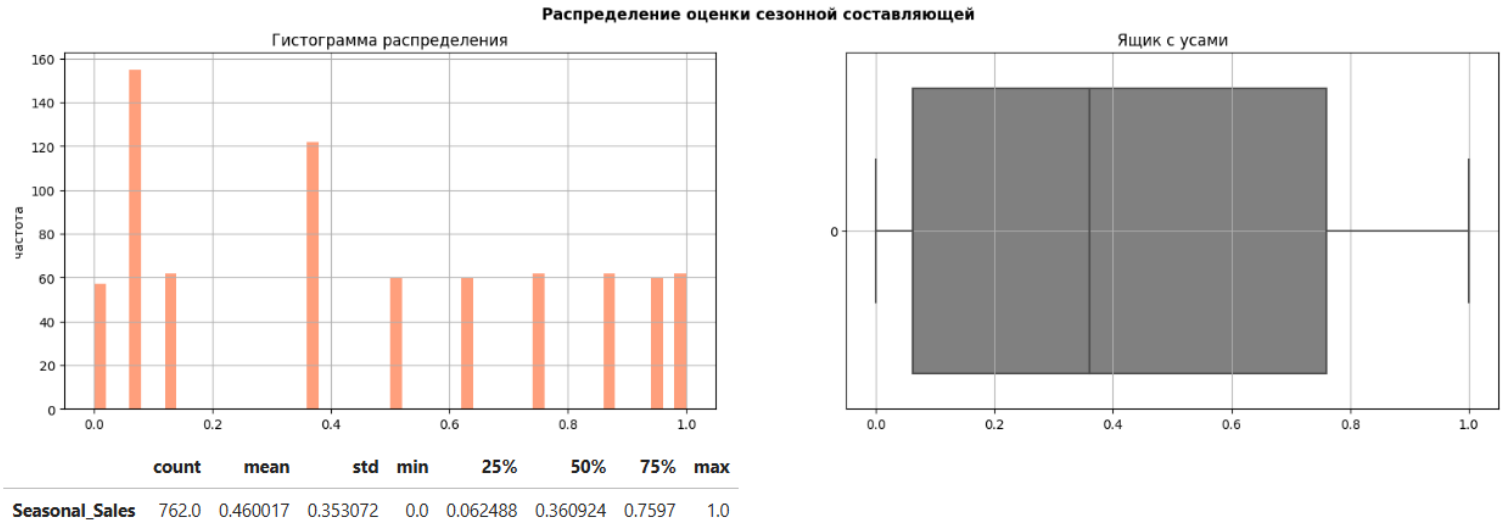
Медианное значение:

2460750 контактов в месяц

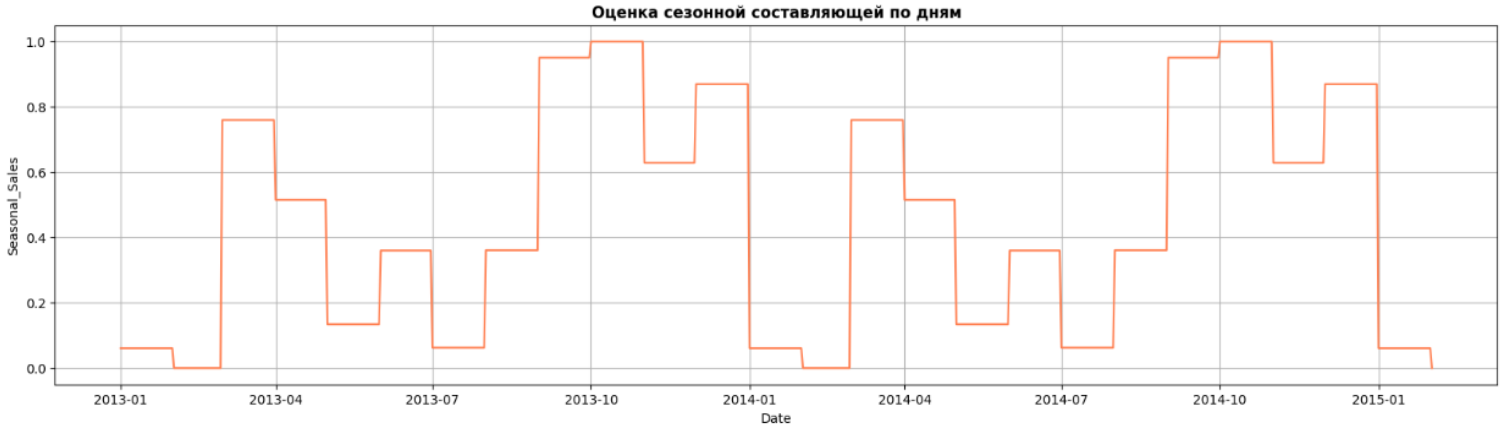
Среднее:

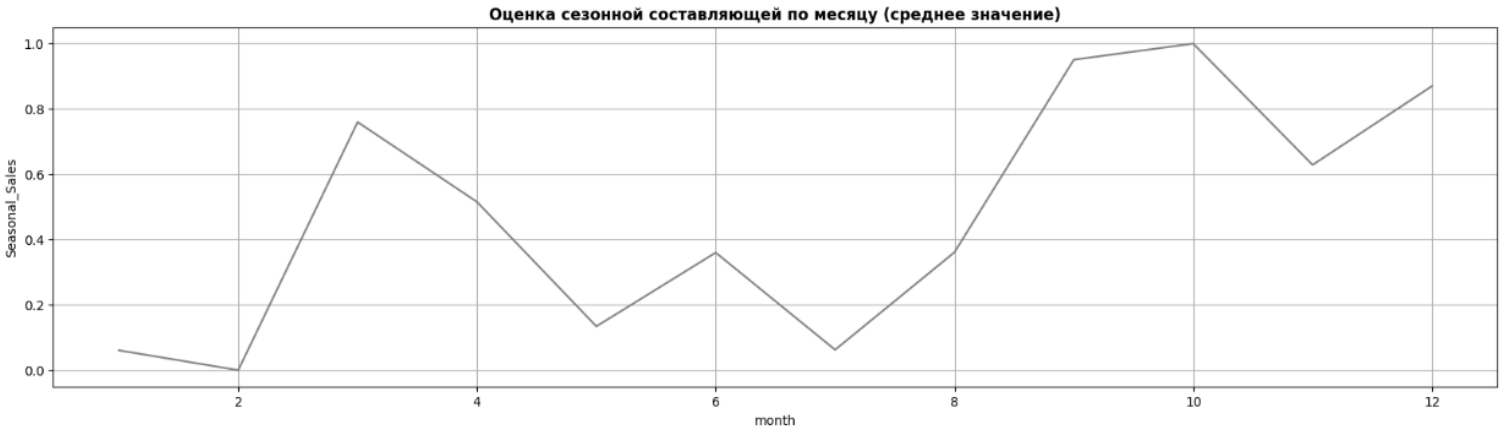
13734923.077 контактов в месяц. Сильно завышается из-за пика в июле 2014

Seasonal_Sales – сезонная составляющая



	count	mean	std	min	25%	50%	75%	max
Seasonal_Sales	762.0	0.460017	0.353072	0.0	0.062488	0.360924	0.7597	1.0





По большей части нормированная оценка сезонной составляющей продаж автомобиля не превышает значения 0.5.

В половине случаев она ниже 0.36.

Максимум:

1

Среднее и медианное:

0.46 и 0.36 соответственно

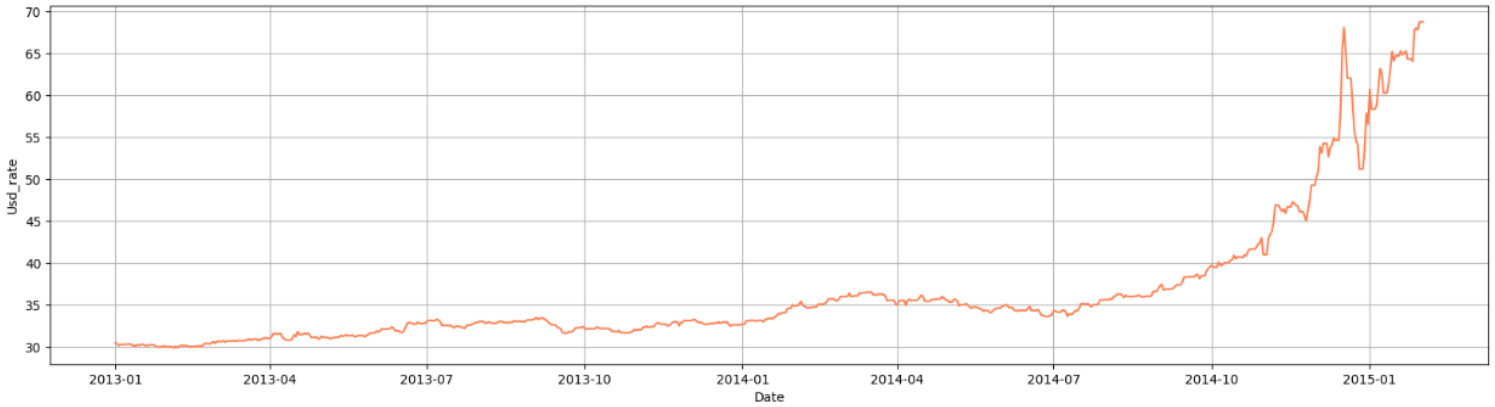
Видим наличие сезонности:

1) низкие значения в январе-феврале
2) скачок в марте
3) снижение к июлю
4) повышение до пикового значения в октябре
5) спад в январе

Доля выше 0.5 в месяцах:

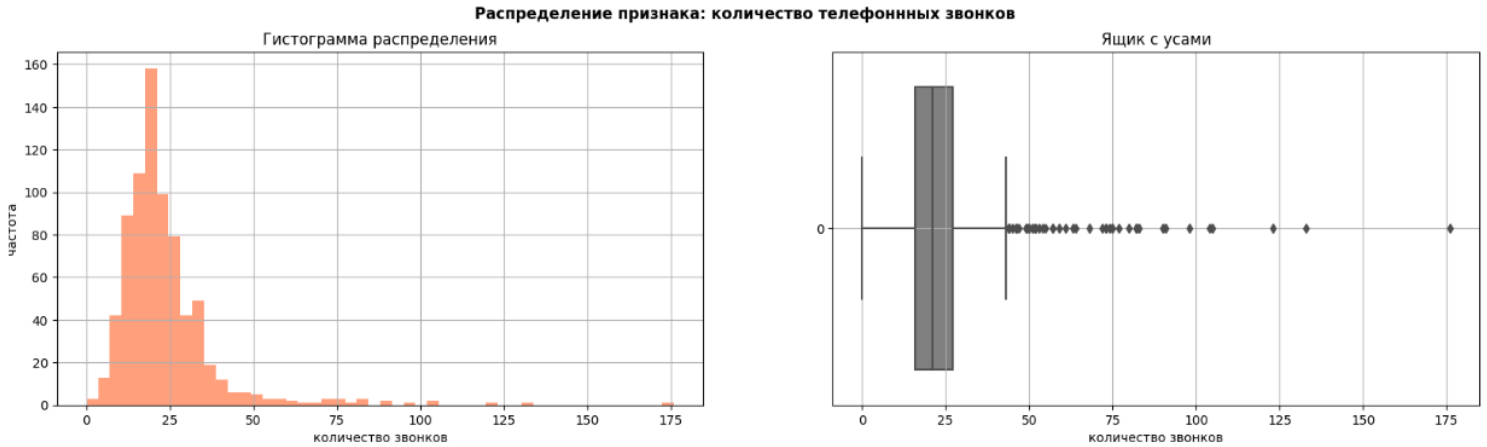
март, апрель, сентябрь, октябрь, ноябрь, декабрь

Usd_rate – курс доллара



- Курс доллара начал явный рост после июля 2014.
- К февралю 2015 вырос примерно в два раза в сравнении с январем 2013.

Y – количество звонков, целевая переменная



	count	mean	std	min	25%	50%	75%	max
Y	759.0	23.627141	15.08155	0.0	16.0	21.0	27.0	176.0

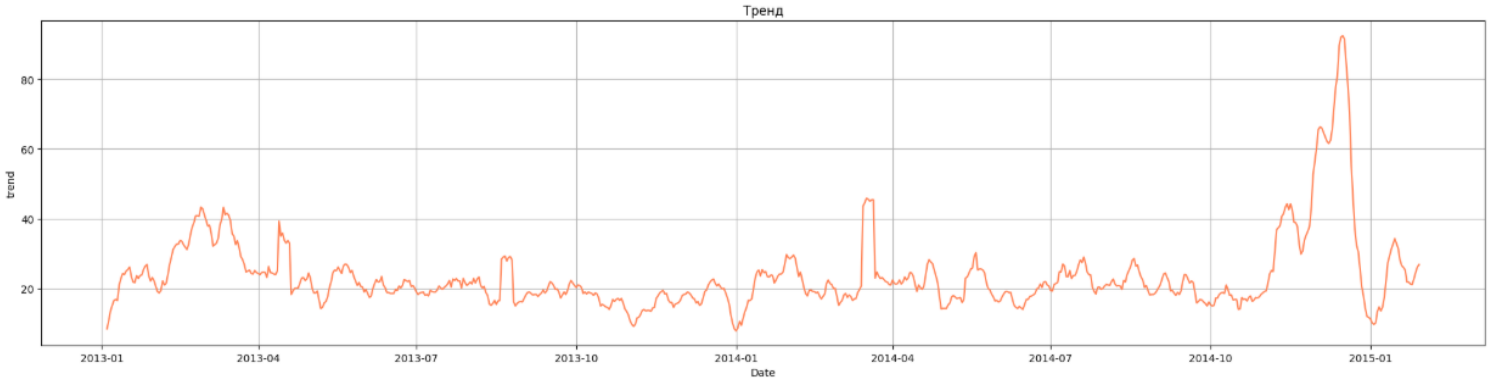
Распределение:

- от 0 до 176 в день
- в среднем - 23.6 в день
- в половине случаев - 16 до 27

На графике видим по крайней мере 3 явных выброса. Это может быть ошибка в данных (например, значения увеличены в 10 раз), но корректировать на всякий случай не будем: данные могут быть реальны и связаны с каким-то другим событием, не указанным в датасете.

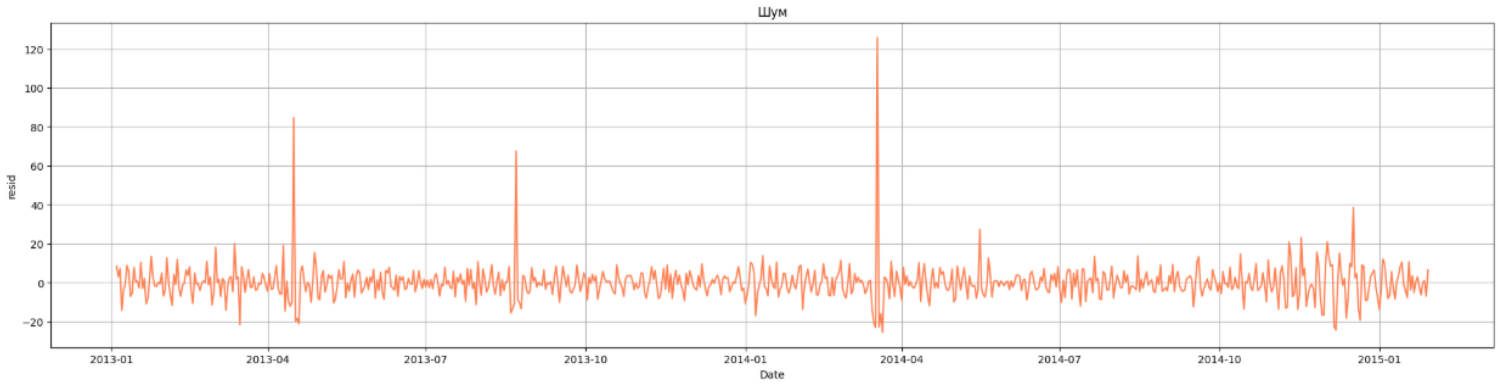
Среднее количество звонков:

- относительно высокое в марте-апреле 2013.
- спад - с апреля 2013 по ноябрь 2013.
- явный быстрый рост - с ноября 2013 по январь 2014.



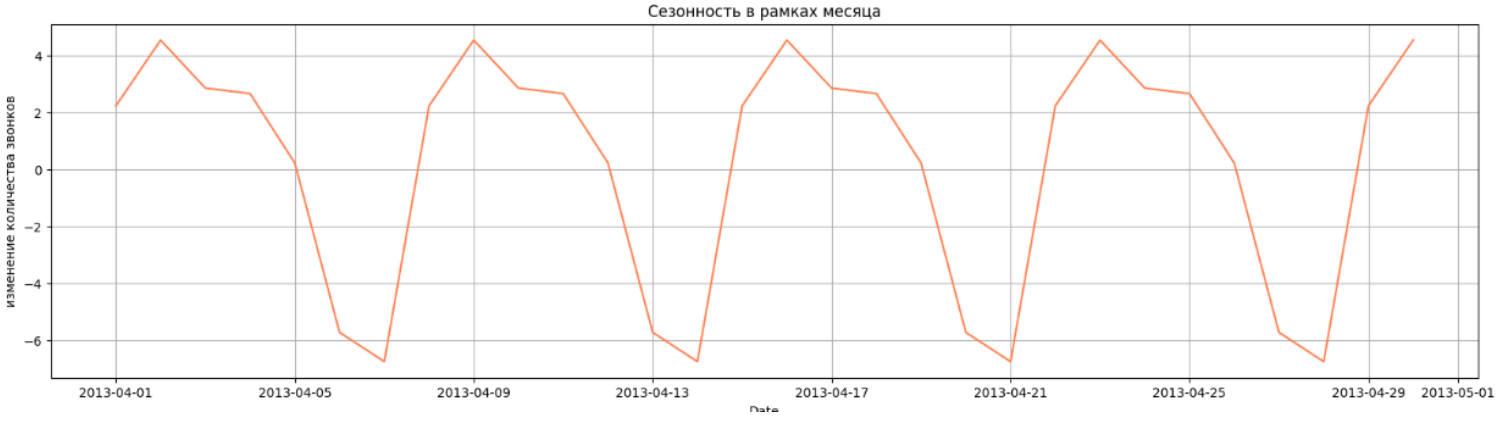
Тренд:

- видим сильный рост к концу 2013 года. Однако, его нельзя назвать устойчивым, т.к. в январе резко пошел на спад.



Шум:

- резкие скачки звонков в апреле и августе 2023, марте 2024. Наблюдается большой разброс в последние месяцы 2014.



Сезонность, месячная:

- С понедельника по четверг включительно - количество звонков наибольшее
- Пятница - спад
- Выходные - низкое количествено звонков

Зависимости целевого признака от входящих

Курс доллара:

- имеет сильную корреляцию со скачком количества звонков в ноябре-декабре 2014.

Наружная реклама:

- Увеличение контактов с аудиторией в марте-апреле 2014 совпадает со скачком количества суммарных звонков в этот же временной промежуток.

Число контактов по ТВ:

- Кажется, слабо влияет на количество звонков. В последней рекламной кампании могло повлиять на рост количества звонков в совокупности с курсом доллара.

Число контактов по ТВ и Наружная реклама:

- в 2 случаях совпадают по времени с увеличением контактов через наружную рекламу. Возможно, их увеличение в марте 2013 повлияло на дальнейший пик в количестве звонков.

Общий вывод по исследовательскому анализу

Данные представлены во временном промежутке 761 дня:

- с 1 января 2013 по 1 февраля 2015

Всего было 3 рекламные кампании по ТВ: 17 февраля - 11 марта 2013 , 23 июня - 13 июля 2014, 24 ноября - 14 декабря. Оценка ежедневного количества с целевой аудиторией по этому каналу в дни кампании – от 5 до 26 млн контактов.

В первой ТВ-кампании корреляция с целевым признаком составила 0.284, во второй – 0.0, в третьей – 0.52.

В 42% месяцев в представленном датасете число контактов по наружной рекламе равно нулю. Всего наружная реклама проводилась в 13 из 25 представленных в датасете месяцев. В среднем – 145.137 млн контактов в месяц. Пики – 2013: март; 2014: март, апрель, июль.

По большей части нормированная оценка сезонной составляющей продаж автомобиля не превышает значения 0.5. В половине случаев она ниже 0.36. Доля выше 0.5 в месяцах: март, апрель, сентябрь, октябрь, ноябрь, декабрь. Видим наличие сезонности: 1) низкие значения в январе-феврале, 2) скачок в марте, 3) снижение к июлю, 4) повышение до пикового значения в октябре, 5) спад в январе.

Курс доллара начал явный рост после июля 2014. К февралю 2015 вырос примерно в два раза в сравнении с январем 2013. С этим ростом связан и сильный рост количества звонков в ноябре-декабре 2014.

На графике целевого признака от даты есть явные 3 явных выброса. Видим сильный рост к концу 2013 года. Однако, его нельзя назвать устойчивым, т.к. в январе резко пошел на спад.

Явно прослеживается сезонность по количеству звонков в течение недели:

> С понедельника по четверг включительно - количество звонков наибольшее. Пятница – спад. Выходные - низкое количество звонков

Взаимосвязи целевого признака со входными:

- Увеличение контактов с аудиторией по наружной рекламе в марте-апреле 2014 совпадает со скачком количества суммарных звонков в этот же временной промежуток.
- Увеличение количества звонков имеет сильную корреляцию со скачком количества звонков в ноябре-декабре 2014.
- Число контактов по ТВ, кажется, слабо влияет на количество звонков. В последней рекламной кампании могло повлиять на рост количества звонков в совокупности с курсом доллара.
- Число контактов по ТВ и Наружная реклама в 2 случаях совпадают по времени с увеличением контактов через наружную рекламу. Возможно, их увеличение в марте 2013 повлияло на дальнейший пик в количестве звонков.

3. Создание новых признаков

Рост курса доллара:

- 'Usd_rate_growth_d' – как изменился курс в сравнении с предыдущим днем.

Отстающие значения:

Y – количество звонков (целевая переменная):

- 'Y_lag_{i}' – 1, 2, 3, 4, 5, 7, 8, 14, 21, 30 дней

TV – контакты через ТВ-рекламу:

- 'TV_lag_{i}' – 1, 2, 3, 5, 7, 14 дней

OOH – контакты через наружную рекламу:

- 'OOH_lag_30' – 21 день

4. Обучение модели – результат

Выбранная метрика:

- MAE – т.к. наименее чувствительна к выбросами и наиболее понятна для интерпретации.

Прим:

- обучение проходило дважды: с учетом трех выбросов, обнаруженных при анализе данных, и без этих выбросов (были заменены на значения недельной давности)

Параметры предобработки и разбиения на выборки:

- размер тестовой выборки – 0.25
- shuffle=False
- масштабирование не применялось

---

1. **С выбросами:**

Обучены модели со следующими метриками на кросс-валидации:

- LinearRegressor:
  - MAE = 7.01
- LightGBM Regressor:
  - MAE = 7.70
- CatBoostRegressor:
  - MAE = 6.64

2. **Без выбросов:**

Обучены модели со следующими метриками на кросс-валидации:

- LinearRegressor:
  - MAE = 5.88
- LightGBM Regressor:
  - MAE = 6.19
- CatBoostRegressor:
  - MAE = 5.85

Лучшая модель:

- CatBoost 2 ('depth': =16, 'iterations': 500, 'l2_leaf_reg': 20, 'learning_rate': 0.6)

---

3. **Результаты на тесте**

MAE = 11.76

---

4. **Проверка на адекватность:**

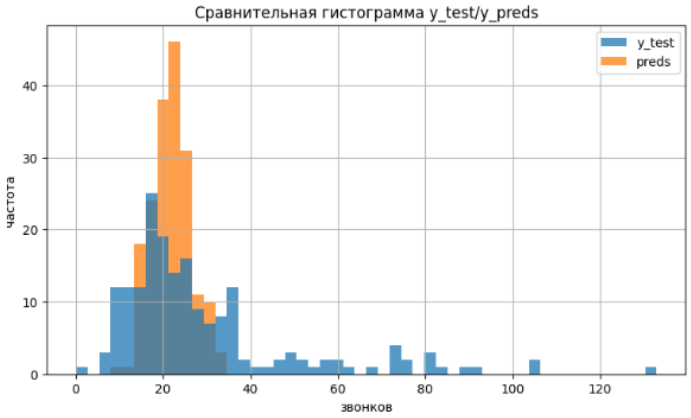
- MAE = 11.76 - предсказания модели
- MAE = 14.97 - предсказания константным средним значением

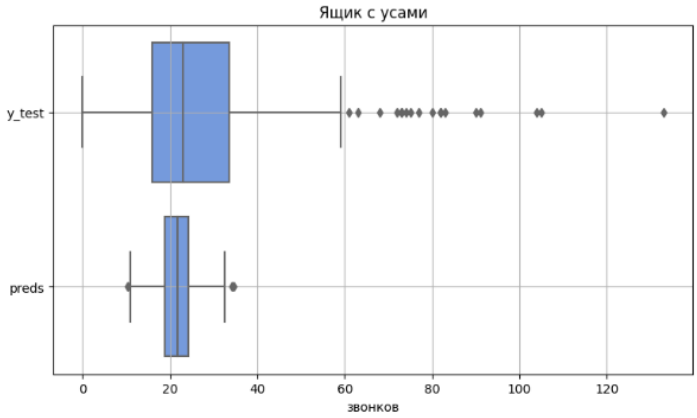
Видим, что модель чуть более эфеективна, чем если бы всегда предсказывали константой - средним.

---

5. Анализ остатков

Сравнительная гистограмма:



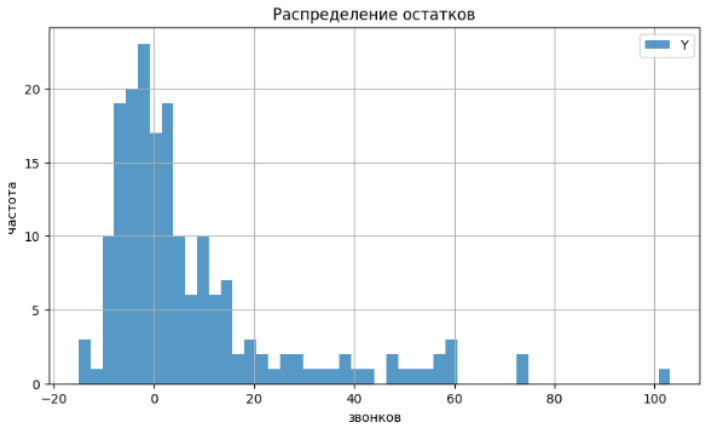


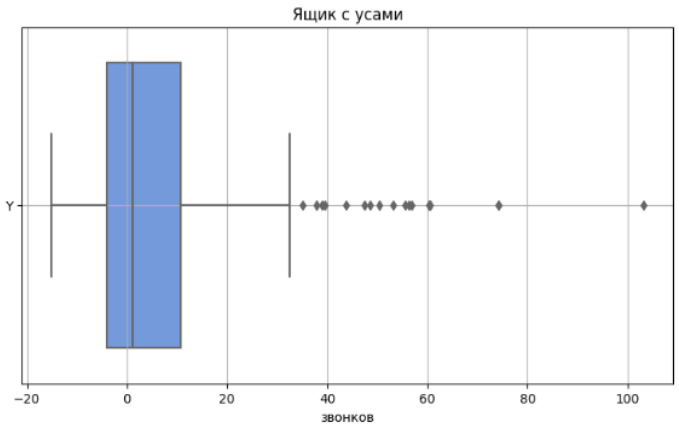
	count	mean	std	min	25%	50%	75%	max
y_test	183.0	29.256831	21.360965	0.000000	16.000000	23.000000	33.500000	133.00000
preds	183.0	21.796037	4.546145	10.390663	18.770941	21.728403	24.283786	34.54089

Распределение значений модели близко к нормальному, в отличие от реальных данных, где значения распределены не очень равномерно.

Заметно ниже стандартное отклонение предсказанных значений: 4.5 против 21.

Распределение остатков:





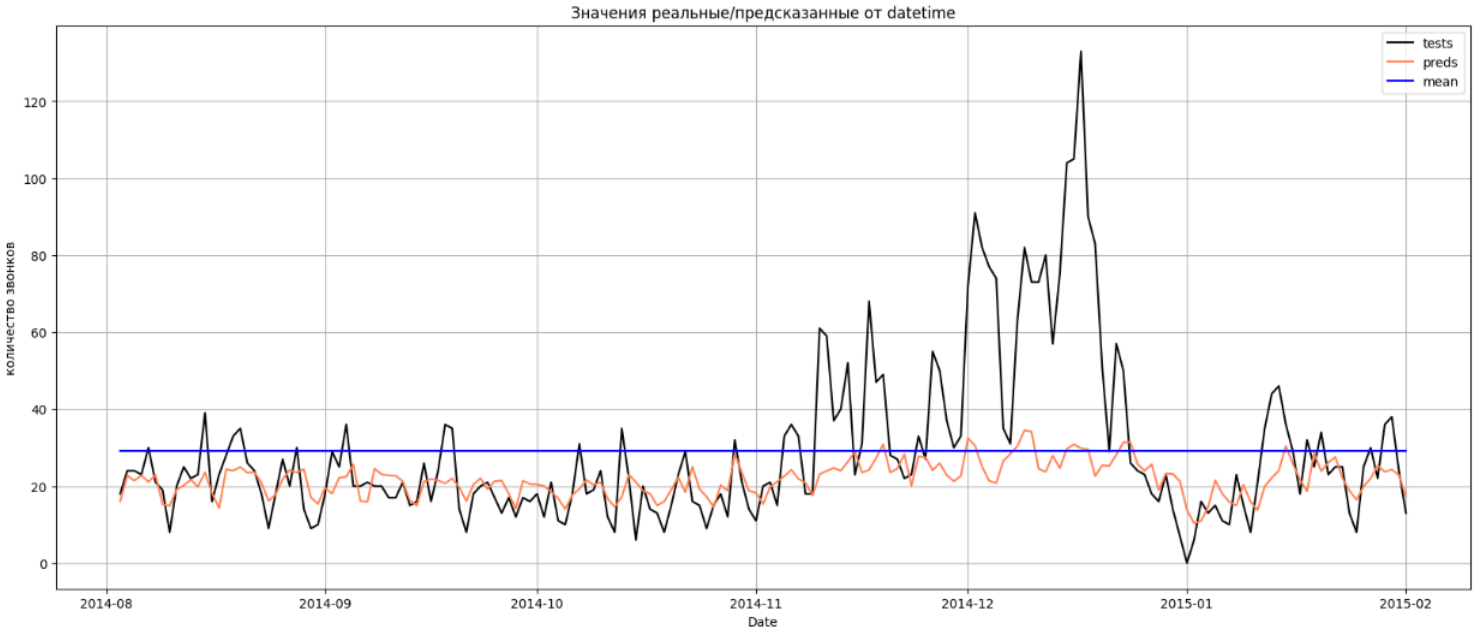
	count	mean	std	min	25%	50%	75%	max
Y	183.0	7.460793	18.826063	-15.044023	-4.115111	1.198929	10.757674	103.172842

Чаще модель преуменьшает значения.

Медиана распределения остатков - 1.2, довольно близко к нулю.
При этом среднее - 7.46, достаточно далеко от нуля.

- увеличение среднего остатков связано с тем, что модель плохо предсказала очень высокие значения декабря 2014 - при обучении она с такими значениями не сталкивалась.

Максимальное завышение количества звонков - 15.

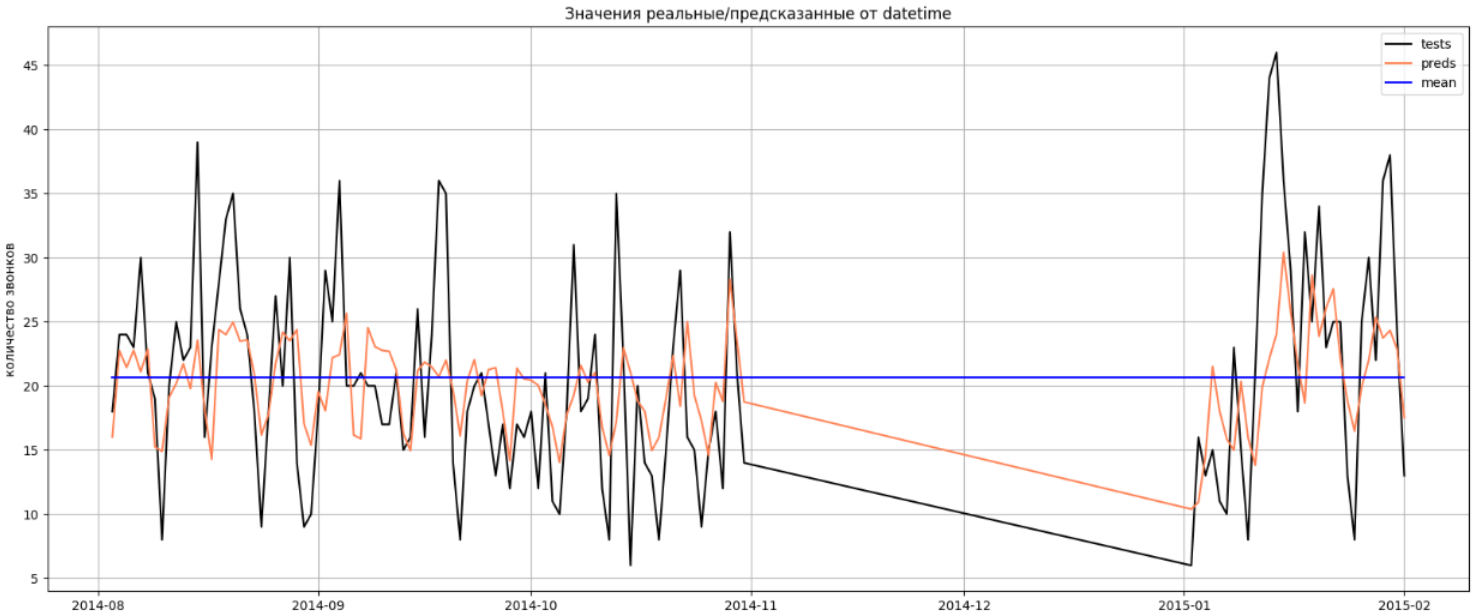


Выводы:

1. Сравнительно неплохо модель предсказывает значения в период до ноября 2014, когда начался резкий скачок звонков.
2. Модель не очень точна, но в основном верно определяет направления изменений количества звонков.
3. Модель не смогла предсказать сильный рост в ноябре-декабре 2014. Скорее всего, именно это значительно ухудшило метрику на тесте.

---

6. Анализ остатков без учета скачка ноябрь-декабрь 2014



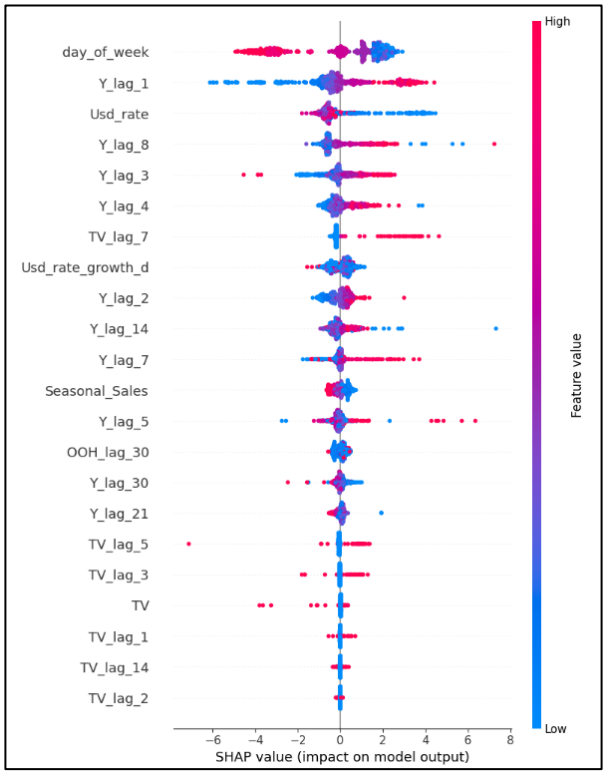
- MAE = 5.61 - предсказания модели без учета ноября-декабря
- MAE = 6.71 - предсказания константным средним значением без учета ноября-декабря

Вывод:

- Метрика значительно выше без учета ноября-декабря - с теми месяцами модель явно не справляется.
- При этом значение метрике все равно не сильне лучше, чем при предсказании константным средним значением.
- Точно можно сказать, что модель неплохо предсказывает направление изменения количества звонков. В большинстве случаев справляется с определением пиковых количеством звонков (наибольших и наименьших).

7. Анализ важности признаков

Наиболее важные с точки зрения модели признаки:



Повышают значение целевой переменной высокие значения:

- количества звонков: вчерашнее, 3, 4, 5, 7, 8 дней назад
- количества контактов по ТВ: 7 дней назад

Повышают значение целевой переменной низкие значения:

- курса доллара
- количества звонков 30 дней назад
- оценки сезонной составляющей продаж

Изменение значения относительно среднего - **повышает** значение целевой переменной:

- TV, TV_lag_1, TV_lag_3, TV_lag_5, TV_lag_14

8. Возможности по улучшению модели

Добавить больше актуальных данных по 2015 году.

Возможно, добавить новый категориальный признак: рост доллара на критический процент (50-100%) в последние день/неделю/месяц/2 месяца . В связи с резким ростом курса доллара в конце года покупательское поведение изменилось – такой признак поможет в дальнейшем учесть подобные изменения. Но для этого также нужно больше данных.

Если есть другие рекламные активности – учесть и их.

С имеющимся набором – можно дополнительно поработать над созданием новых признаков. Подобрать больше отстающих значений, скользящих средних.

Также количество отстающих значений можно нарастить до 365 – при этом удалится значительная часть датасета, но останутся данные по целому году и, возможно, модель сможет учесть какую-то сезонность по году (например, первые дни января всегда проседают по звонкам). Скорее всего, Y_log_365 может стать одним из ключевых признаков для модели.

Возможно, из ООН и TV тоже можно было бы попробовать сделать категориальный признак – идет ли реклама в текущем месяце.