1. Сколько магазинов формата стрит в тестовой выборке?

4 магазина формата ‘Стрит’, 1 с формата ‘Стрт’, 18 формата Street (где все буквы латинские), формата Stееt (ее – буквы кириллицы). В сумме 24 магазина формата стрит.

1. Какова средняя выручка магазинов формата мини ТЦ за 2016 год?

307234.8478048167

1. Сколько магазинов с бесплатной парковкой?

156 магазин с бесплатная парковка и 1 бесплатная паpковка (буква ‘o’ из латиници).

Замечания: Много категориальных признаков, которые требуют предобработку перед загрузкой в систему.

Есть признаки, интерпретация которых не понятна, например признаки 'Расположение в ТЦ', 'Конкуренты в ТЦ', 'Рядом с эскалатором' и 'Рядом со входом'. Для 'Рядом с эскалатором' и 'Рядом со входом' лучше подойдут не дискретные значения, а лучше вещественные и иметь представление что они значат. Так как эти дискретные значения могут означать номер этажа от входа или по какой-то шкале как далеко находиться магазин.

Так же можно проверить какой импакт вносят эти признаки в обученной модели.

Признаки 'Формат магазина' и 'Парковка' в значениях имеют опечатки в значениях:

{'Street', 'Strееt', 'Мини ТЦ', 'Мини-ТЦ', 'Стрит', 'Стрт', 'ТЦ'}

{100,

1000,

700,

nan,

'бесплатная паpковка',

'бесплатная парковка',

'нет парковки',

'платная парковка'}

При этом визуально они написаны одинаково, хотя во втором Street обе ‘ee’ являются русскими, похожая ситуация с бесплатная парковка в признаке 'Парковка'.

А также не самые понятные значения 100, 1000, 700. Возможно, это стоимость парковки, но не ясно за какое время и какая парковка.

Признак 'Время работы' в своих значениях имеет одинаковые промежутки работы магазина, но в них допущены опечатки и из-за этого появляются копии значений в которых нет необходимости.

Северная точка — 81°50′35″ с. ш. 59°14′22″ в. д.

Южная точка — 41°11′07″ с. ш. 47°46′55″ в. д.

Западная точка — 54°27′45″ с. ш. 19°38′19″ в. д.

Восточная точка — 65°47′ с. ш. 169°01′ з. д.

Широта и долгота расположена почти в пределах России, но широта и долгота перепутаны места если сравнить данные из википедии о крайних точках России.

91.75824899999999 – максимальная долгота, которая должна была являться широтой.

Долгота == широте, соответственно, самый крайний по долготе магазин находиться за пределами полярного круга. Значение от экватора может равняться только от 0 до 90 градусов.

Из вышеизложенного предложения следует, что необходимо проверить адреса магазинов, возможно некоторых из них просто не существует, но для этого нужно написать скрип и не проверять руками.

Дата закрытия магазинов вполне валидная, но не понятно зачем для них нужно делать анализ о прогнозировании выручки.

Но дата открытия не удовлетворяет условию в письме, где было сказано, что данные о выручке с 2015 года. Потому что есть выбросы с датой открытия 2098 года, но скорее всего это опечатка.

Признак 'Площадь магазина' имеет странные значения, которые вызывают вопросы. Наверняка есть магазины с такой большой площадь, но значение 1488 требует отдельного внимания. Есть факт, что магазин с такой площадью существует или это клиент сделал специально и через данные что-то хотел сказать?

Признак 'Площадь ТЦ' выглядит вполне нормально, кроме пропусков. Необходимо узнать у клиента есть ли возможность дополнить эту недостающую информацию.

Признак 'Этаж' имеет некорректные значения в виде даты

Признак 'Заполненность ТЦ', имеет 26% пропущенных значений. Стоит спросить появиться ли информация о заполненности.

Распределение признака 'Толока' похоже на Пуассоновское. Я не совсем понимаю, что может означать этот признак, возможно это коэффициент работников магазина, либо значение того, как покупатели делают покупки.

Признак 'Люкс' выглядит вполне нормально. Но тоже сложно интерпретировать эти значения. Люкс — это значение насколько цены в магазине выше из-за наценки за бренды, которые там продаются, либо наценки за качество.

Главный вопрос из-за количества объектов в выборке. Точно ли данной компании необходимо МЛ решение, потому что-то не похоже, что у них имеется достаточное количество данных для обучения моделей. Я понимаю, что этот дата сет только для тестового задания)