**Кейс «Моделирование пиковых часов потребления электроэнергии»**

1. **Целевые переменные**

**В качестве целевых переменных для построения интервала из 4-х прогнозных часов предлагается:**

1. **для задачи классификации:**
2. ***max\_ec* – отметка часа, в котором потребление энергии (*energy\_consumption*) было максимальным. Принимает значения [0,1]. Один флаг внутри одного дня.**
3. ***new\_class* – количество интервалов для данного часа, которые могли бы быть верными при расчёте итоговой метрики. Внутри дня принимает значения [4,3,2,1,0] в количестве [1, 2, 2, 2, 17] соответственно.**
4. ***best\_interval\_flag* – флаг на каждый из 4х часов с наилучшим для прогнозирования интервалом. Интервал определяется как 4х часовой интервал с максимальным потреблением энергии, включающий в себя час с фактическим максимальным потреблением энергии. Принимает значения [0,1]. Четыре флага внутри одного дня.**
5. **для задачи регрессии:**
6. ***energy\_consumption* – потребление энергии в данный час**
7. ***yms\_ consumption\_minmax* – нормализованное по *MinMax* значение energy\_consumption внутри одного дня. Принимает непрерывные значения в интервале от 0 до 1.**
8. ***rank\_day* – присвоенный ранг для каждого часа по возрастанию после ранжирования по потреблению энергии energy\_consumption. Принимает целочисленные значения от 0 до 23.**
9. **Постобработка прогнозов для расчёта метрики.**

**Прогнозный интервал определялся как интервал, в котором сумма предсказанных вероятностей или абсолютных значений для четырёх соседних часов была максимальной. (По аналогии с расчётом предсказаний базовой модели)**

1. **Undersampling.**

**Для увеличения доли позитивных классов и повышения эффективности моделей машинного обучения был применён *undersampling*: из обучающей выборки удалены крайние значения часов, в которые ни разу не попадали пиковые значения часов: часы от 0 до 6 включительно и 21 до 23 включительно.**

**На 6-ой час пришлось 3 фактических максимальных часа, но они не попали в строгий интервал плановых пиковых часов. Данные три дня были предобработаны.**

1. **Feature engineering.**

**Кроме временных признаков изначально в обучающей выборке нет признаков, которые можно было бы использовать для обучения, т.к. они все несут в себе информацию из будущего. Были созданы новые признаки путём преобразований, группировок и сдвигов различных признаков во времени.**

**Помимо основного временного ряда с шагом в 1 месяц внутри дня присутствует временной ряд с шагом в 1 час. Были сформированы новые признаки описывающие данный временной ряд.**

**При увеличении количества признаков растёт время обучения модели, поэтому количество сгенерированных признаков было ограничено.**

**В качестве новых признаков были использованы:**

1. **предсказания статистических моделей:**
   1. **baseline(базовая модель)**
   2. **stat\_model\_1**
   3. **stat\_model\_2**
   4. **stat\_model\_3**
2. **сдвиги и преобразования во временном ряду различных признаков:**
   1. **сдвиг на 1, 6 и 12 месяцев**
   2. **среднее за 3, 6, 12 предыдущих месяцев**
   3. **разницу между средними:**
      1. **прошлого месяца с позапрошлым**
      2. **прошлого месяца со средним за последние 3 месяца**
      3. **прошлого месяца с месяцем с лагом в год**

**В общей сложности сгенерировано приблизительно 600 новых признаков.**

1. **Feature selection.**

**Отбор признаков для построения моделей проводился с использованием алгоритмов LightGBM, RandomForest и их встроенных методов ранжирования важности признаков. Также во внимание принималась визуальная оценка динамики изменения метрики во время отбрасывания признаков.**

1. **Cross-validation, testing.**

**После генерации признаков на временных рядах из выборки был отброшен год. Выборка стала представлена временным интервалом с сентября 2020 года до октября 2022 года включительно.**

**Обучающая выборка была условно разделена на две части:**

1. **Валидационная часть, которая состоит из 20 месяцев, для отбора признаков, подбора гиперпараметров, кросс-валидации. Данная часть была условно разделена на:**
   * + 1. **12 месяцев для первоначального обучения**
       2. **8 месяцев для последовательного прогнозирования с шагом в один месяц**

**2. Тестовая часть для оценки моделей, которая состоит из 6 месяцев, начинается с мая 2022г и заканчивается октябрём 2022г.**

**Кросс-валидация и тестирование проводится с шагом в 1 месяц.**

**Все этапы валидации и обучения проводились на выборке, в которой совместно использовалась информация и из Пермского края, и из Тюменской области.**

1. **Выбор метрик.**

**Для оценки итоговой эффективности моделей использовалась предложенная в задании метрика, которая определяется как доля верно найденных пиков дня при использовании 4х часового интервала.**

**Для отбора признаков, подбора гиперпараметров и дополнительной оценки эффективности модели использовалась метрика Roc-auc, которая условно показывает эффективность ранжирования вероятностей выпадения классов. А также MAE, которая определяется средней ошибкой предсказания в задачах регрессии.**

1. **Подбор гиперпараметров.**

**Подбор гиперпараметров проводился после отбора признаков только для моделей, которые основаны на градиентном бустинге. Для оценки эффективности подбора использовалась метрика Roc-Auc для классификации и MAE для регрессии.**

1. **Выбор алгоритмов машинного обучения.**

**В качестве простых базовых решений использованы несколько вариантов статистических моделей, которые строились по принципам базовой модели.**

**Для объединения предсказаний статистических моделей и представления быстрого решения машинного обучения выбрана линейная модель LogisticRegression**

**Для формирования более сложного решения был выбран градиентный бустинг LightGBM, который, исходя из личного опыта, отличается своей скоростью обучения в сравнении с бустингами CatBoost и XGBoost.**

1. **Описание моделей.**

**baseline (базовая модель) называет интервал, в котором было больше всего пиков в прошлом году в том же месяце. *(max\_ex)***

**stat\_model\_1 называет интервал, в котором было больше всего попадания в верный интервал в прошлом году в том же месяце. *(new\_class)***

**stat\_model\_2 называет интервал, в котором было больше всего пиков в прошлом году в том же месяце в то же время недели (начало недели: понедельник - вторник, середина: среда-четверг, конец: пятница – суббота). *(weekday, max\_ec)***

**stat\_model\_3 называет интервал, в котором было больше всего попаданий в верные интервалы в прошлом году в том же месяце в то же время месяца (первая половина месяца: до 17-го числа, вторая половина месяца: после 17го числа включительно) *(half\_month, new\_class)***

**Описание моделей с использованием машинного обучения**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Модель** | **Целевая переменная** | **Признаки** | **Отбор признаков** | **Гиперпар-ры** |
| LogisticRegression | **Cl: max\_ec** | **временные признаки + предсказания стат. моделей** | **нет** | **нет** |
| LGBMClassifier  \_max\_ec | **Cl: max\_ec** | **временные признаки + предсказания стат. моделей + ~600 признаков на временных рядах** | **нет** | **нет** |
| LGBMClassifier  \_best\_interval\_flag | **Cl: new\_class** | **нет** | **нет** |
| LGBMClassifier  \_new\_class | **Cl: best\_interval\_flag** | **нет** | **нет** |
| LGBMClassifier  \_feature\_selected  \_best\_interval\_flag | **Cl: best\_interval\_flag** | **~ 150 отобранных признаков** | **да** | **нет** |
| LGBMClassifier  \_feature\_selected  \_hyperparams  \_best\_interval\_flag | **Cl: best\_interval\_flag** | **~ 150 отобранных признаков** | **да** | **max\_depth = 5,**  **num\_leaves = 37** |
| LGBMRegressor  \_energy\_consumption | **Reg: energy consumption** | **временные признаки + предсказания стат. моделей + ~600 признаков на временных рядах** | **нет** | **нет** |
| LGBMRegressor  \_rank\_day | **Reg: yms\_energy\_consumption\_minmax** | **нет** | **нет** |
| LGBMRegressor  \_ymds\_energy  \_consumption\_minmax | **Reg: rank\_day** | **нет** | **нет** |
| LGBMRegressor  \_feature\_selected  \_rank\_day | **Reg: rank\_day** | **63 отобранных признаков** | **да** | **нет** |
| LGBMRegressor  \_feature\_selected  \_hyperparams\_rank\_day | **Reg: rank\_day** | **63 отобранных признаков** | **да** | **max\_depth = 6,**  **num\_leaves = 24** |

1. **Принципы оценки модели.**

**Предлагается сравнивать модели по следующим критериям:**

1. **Основной критерий:**
2. **Основная метрика из условия задачи на тестовой выборке после прохождения предварительной валидации**
3. **Дополнительные критерии**
4. **Метрика Roc-Аuc / MAE (сравнивать только между моделями, которые обучены на одну и ту же целевую переменную)**
5. **Основная метрика на валидации**
6. **Стандартное отклонение или другой показатель изменчивости основной метрики по месяцам на тесте и кросс-валидации**
7. **Разница между результатом на кросс-валидации и тренировочной выборке.**
8. **Сложность в реализации и поддержке моделей.**
9. **Результаты**

**Обобщённые результаты моделей на CV и test**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Модель | cр.Q CV  (8 месяцев) | стд.откл. CV  (8 месяцев) | cр.Q test  (6 месяцев) | стд.откл. test  (6 месяцев) | Разница метрик | Сложность модели |
| baseline | 0,681 | 0,199 | 0,605 | 0,050 | -0,076 | 0 |
| stat\_model\_1 | 0,704 | 0,193 | 0,607 | 0,073 | -0,097 | 0 |
| stat\_model\_2 | 0,663 | 0,162 | 0,612 | 0,044 | -0,051 | 0 |
| stat\_model\_3 | 0,699 | 0,160 | 0,596 | 0,063 | -0,103 | 0 |
| LogisticRegression | 0,688 | 0,197 | 0,663 | 0,053 | -0,024 | 1 |
| LGBMClassifier\_max\_ec | 0,688 | 0,162 | 0,637 | 0,133 | -0,051 | 2 |
| LGBMClassifier\_best\_interval\_flag | 0,715 | 0,160 | 0,697 | 0,077 | -0,018 | 2 |
| LGBMClassifier\_new\_class | 0,701 | 0,141 | 0,665 | 0,126 | -0,036 | 2 |
| LGBMClassifier\_feature\_selected\_  best\_interval\_flag | 0,709 | 0,166 | 0,735 | 0,096 | 0,027 | 2 |
| LGBMClassifier\_feature\_selected\_  hyperparams\_best\_interval\_flag | 0,711 | 0,159 | 0,737 | 0,088 | 0,026 | 2 |
| LGBMRegressor\_  energy\_consumption | 0,598 | 0,222 | 0,618 | 0,109 | +0,019 | 2 |
| LGBMRegressor\_rank\_day | 0,612 | 0,228 | 0,672 | 0,081 | +0,061 | 2 |
| LGBMRegressor\_ymds\_energy\_  consumption\_minmax | 0,580 | 0,241 | 0,653 | 0,063 | +0,073 | 2 |
| LGBMRegressor\_feature\_selected\_  rank\_day | 0,613 | 0,221 | 0,692 | 0,087 | +0,078 | 2 |
| LGBMRegressor\_feature\_selected\_  hyperparams\_rank\_day | 0,623 | 0,212 | 0,704 | 0,109 | +0,081 | 2 |

**Сравнение результата с базовой моделью**

|  |  |  |  |
| --- | --- | --- | --- |
| Год-месяц | LogisticRegression | LGBMСfshbif\* | LGBMСfshrd\*\* |
| 2022-05 | 0,000 | 0,056 | -0,028 |
| 2022-06 | 0,119 | 0,095 | 0,119 |
| 2022-07 | 0,143 | 0,167 | 0,071 |
| 2022-08 | 0,087 | 0,043 | 0,000 |
| 2022-09 | 0,000 | 0,205 | 0,227 |
| 2022-10 | 0,000 | 0,225 | 0,200 |
| В среднем | +5,8% | + 13,2% | +9,8% |

\*LGBMClassifier\_feature\_selected\_hyperparams\_best\_interval\_flag

\*\*LGBMClassifier\_feature\_selected\_hyperparams\_rank\_day

**Сравнение результата с базовой моделью для Пермского края**

|  |  |  |  |
| --- | --- | --- | --- |
| Год-месяц | LogisticRegression | LGBMСfshbif\* | LGBMСfshrd\*\* |
| 2022-05 | 0,000 | 0,111 | -0,056 |
| 2022-06 | 0,000 | -0,048 | 0,000 |
| 2022-07 | 0,143 | 0,190 | 0,143 |
| 2022-08 | 0,000 | -0,087 | 0,000 |
| 2022-09 | 0,000 | 0,409 | 0,455 |
| 2022-10 | 0,000 | 0,450 | 0,400 |
| В среднем | + 2,4 | +17,1% | +15,7% |

\*LGBMClassifier\_feature\_selected\_hyperparams\_best\_interval\_flag

\*\*LGBMClassifier\_feature\_selected\_hyperparams\_rank\_day

**Сравнение результата с базовой моделью для Тюменской области**

|  |  |  |  |
| --- | --- | --- | --- |
| Год-месяц | LogisticRegression | LGBMСfshbif\* | LGBMСfshrd\* |
| 2022-05 | 0,000 | 0,000 | 0,000 |
| 2022-06 | 0,238 | 0,238 | 0,238 |
| 2022-07 | 0,143 | 0,143 | 0,000 |
| 2022-08 | 0,174 | 0,174 | 0,000 |
| 2022-09 | 0,000 | 0,000 | 0,000 |
| 2022-10 | 0,000 | 0,000 | 0,000 |
| В среднем | +9,2% | +9,2% | +4,0% |

\*LGBMClassifier\_feature\_selected\_hyperparams\_best\_interval\_flag

\*\*LGBMClassifier\_feature\_selected\_hyperparams\_rank\_day

1. **Выводы**

**Среди всех моделей своей эффективностью выделяются модели, которые решали задачу бинарной классификациицелевой переменной *best\_interval\_flag*. По основной метрике лучшей является модель *LGBMClassifier\_feature\_selected\_hyperparams\_best\_interval\_flag,* которая в среднем превышает эффективность базовой на 13%. Для предсказаний модель использует ~150 отобранных признаков. Отбор признаков существенно увеличил метрику и уменьшил риск переобучения. Также для этой модели был проведён неглубокий подбор гиперпараметров.**

**Переход от поиска единственного максимального часа *max\_ec* к мультиклассификации *new\_class* увеличила эффективность модели, которая основана на бустинге.**

**В задаче регрессии переход от прогнозирования абсолютных значений** *energy\_consumption* **к прогнозированию относительных внутридневных значений *rank\_day* и** *ymds\_energy\_consumption\_minmax* **позволил увеличить эффективность модели.**

**Несмотря на то, что регрессионные модели в данном кейсе выглядят слабее классификационных, можно заметить, что они показывают себя хуже на валидации, но имеют сравнительно высокую эффективность на тесте. Возможно, для раскрытия потенциала данных моделей необходимо провести дополнительный разведочный анализ, провести генерацию признаков, корректировку целевой переменной.**

**Логистическая регрессия, которая объединяет преобразованные прогнозы статистических моделей, может использоваться как простой вариант предсказаний, который ощутимо превышает эффективность отдельно взятой базовой и других рассмотренных статистических моделей.**

**Модели, которые в среднем лучше базовой модели в общем случае, в среднем лучше проявляют себя и в отдельных регионах. Драйверами роста эффективности моделей в данном кейсе стал поиск правильной целевой переменной, преобразование существующих признаков и генерация новых признаков на временных рядах.**