УДК 004.855.5

Е.И. Цыбуля, А.И. Гуда

# ПРИМЕНЕНИЕ МЕТОДОВ MACHINE LEARNING К ЗАДАЧЕ ОПРЕДЕЛЕНИЯ ГОТОВНОСТИ РАСПЛАВА МЕТАЛЛА

*Аннотация. Разработана модель прогнозирования готовности расплава электростали по химическому составу с применение методов машинного обучения с учителем.*

*Ключевые слова. Машинное обучение, классификация, модель, предикторы, деревья решений, расплав.*

При выплавке стали в электропечи важными факторами контроля являются содержание элементов в жидком металле и время необходимое для доведения металла до нужного химического состава (готовности к выпуску металла из печи). Важность химического состава расплава обуславливается его влиянием на эксплуатационные свойства металла, а время плавки взаимосвязано со скоростью протекания процессов рафинирования и легирования расплава. Таким образом, необходимость контроля содержания химических элементов в жидкой стали направлено на соблюдение требований нормативной документации по качеству стали заданной марки. В свою очередь экономически обоснованным является определение минимально достаточного времени плавки для каждой марки стали в условиях определенного агрегата и изначального состава загружаемых материалов.

В условиях необходимости постоянного мониторинга химического состава выплавляемой стали становится весьма актуальной задача уменьшения количества процедур отбора проб при сохранении доли доведенных по химическому составу расплавов. Сокращение количества таких процедур на плавку снизит время на получение расплава требуемого состава и уменьшит расходы на проведение анализа химического состава в лабораториях предприятия.

В настоящее время решение о проверке химического состава во время проведения плавки основывается в большинстве случаев на опыте оператора, руководящего процессом получения стали. Таким

образом, фактически применяется экспертная оценка параметров плавки и происходит выбор оператором времени взятия пробы, для которого он руководствуется личными знаниями, опытом и интуицией. Поскольку количество параметров плавки достаточно велико и не связано с конечным химическим составом продукта простыми линейными взаимосвязями, даже эксперту в данной области обычно сложно определить момент наступления соответствия необходимого химического состава находящемуся в данный момент в печи. Следует отметить, что встречаются ситуации (чаще в технической сфере), когда наилучшими являются решения, принятые не на основании накопленного опыта, а вопреки ему.

Решения, принятые на основании экспертной оценки также имеет существенный недостаток, связанный с так называемым «человеческим фактором», что обычно негативно сказывается на целесообразности и точности принятия решений. Кроме того, процесс выплавки стали отличается высокой динамичностью, а также непрерывным изменением и неполнотой данных, и ошибки персонала на производстве могут стоить на только времени на исправление ситуации, но и крупных экономических потерь.

Действенным способом повышения эффективности и качества управления в условиях высокой неопределенности процесса является применение методологии системного анализа и принятие решений на основе математических моделей. Полезность и преимущество методов математического моделирования и статистики в настоящее время уже успели оценить на современных украинских заводах. Так в условиях ЧАО «Днепроспецсталь» были проанализированы массивы из плавок коррозиестойких сталей (КСС), в результате чего установлено, что в определенных соотношениях химического состава металла гарантированно обеспечиваются нормативные требования механических свойств стали, а также технологическая пластичность во время проката и число скручиваний на горячее кручение [1].

Самым перспективным направлением в области математического моделирования в последние годы остается применение методов machine learning (машинного обучения) для задач прогнозирования (классификации или регрессии) [2]. Применение моделей машинного обучения основано на выборе оптимального решения, которое максимизирует (или минимизирует) целевую функцию, моделирующую

степень предпочтительности в направлении достижения цели моделирования. Модели машинного обучения могут применяться во всех областях, где возможно получение разнообразных параметров (предикторов), характеризующих процесс, а также накопление крупных массивов экспериментальных данных. Одной из таких областей является металлургия, где процессы характеризуются высокой степенью неопределенности и трудно поддаются прогнозированию, хотя зачастую агрегаты обеспечены большим количеством датчиков для измерения параметров.

В данной статье мы рассмотрим применение методов машинного обучения с учителем для прогнозирования доведения химического состава при выплавке одной из марок электростали в ДСП.

Для обработки исходных данных и создания модели был выбран язык Python, так как он содержит множество разнообразных инструментов для моделирования и внедрения задач машинного обучения.

В качестве исходных данных были использованы 41 параметр процесса при каждой из 105 плавок одной из марок электростали, среди которых время начала плавки, время взятия пробы, химический состав загружаемого материала, параметры оборудования для осуществления выплавки стали и другие показатели. При этом для использования в модели оба параметра времени были преобразованы из типа Timestamp в единый параметр Time (количество часов после начала плавления) с типом Float. Целевой функцией при построении модели выступал бинарный признак готовности расплава по химическому составу, который определялся на основании взятия пробы в определенный момент времени. Фрагмент получившегося датафрейма с исходными данными представлен на рис. 1.

| | Ind | Time | Electr | Nsr | Wslit | UdN | Badya | Lom4SH | Lom3SH | Lom15 | ... | FeSiMn | Al160 | FeSi65 | DoloSiroj | BKKP4 | Olive | Boloto | VesShlak | Targ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1135296 | 0.858 | 58.6 | 103.6 | 162.6 | 360.4 | 2 | 151.9 | 6.0 | 0.0 | ... | 1303.0 | 150.0 | 600 | 99 | 0 | 180.0 | 12.97 | 16.91 | 0 |
| 1 | 1135296 | 0.558 | 58.6 | 103.6 | 162.6 | 360.4 | 2 | 151.9 | 6.0 | 0.0 | ... | 1303.0 | 150.0 | 600 | 99 | 0 | 180.0 | 12.97 | 16.91 | 1 |
| 2 | 1135297 | 1.169 | 62.1 | 109.2 | 162.2 | 383.0 | 2 | 117.5 | 5.6 | 0.7 | ... | 1298.5 | 166.5 | 552 | 97 | 0 | 112.5 | 13.90 | 8.20 | 0 |
| 3 | 1135297 | 1.164 | 62.1 | 109.2 | 162.2 | 383.0 | 2 | 117.5 | 5.6 | 0.7 | ... | 1298.5 | 166.5 | 552 | 97 | 0 | 112.5 | 13.90 | 8.20 | 1 |
| 4 | 1135299 | 1.705 | 60.6 | 107.4 | 159.4 | 380.2 | 2 | 133.7 | 6.0 | 0.0 | ... | 1299.0 | 0.0 | 604 | 101 | 0 | 0.0 | 34.06 | 7.67 | 0 |

Рисунок 1 - Фрагмент датафрейма с исходными данными

В первом столбце датафрейма также указан id плавок, во время которых осуществлялись заборы проб. Целевая функция бинарного типа и в датафрейме представлена в столбце Targ (здесь 0 — расплав не достиг требуемого химического состава, 1 — достиг).

Подготовка исходных данных также включала в себя проверку на отсутствующие значения в выборке и замена их на 0 при необходимости, а также проверка уровня корреляции между всеми рассматриваемыми предикторами. Проверка корреляций осуществлялась согласно следующему алгоритму (рис. 2).

```
CorrKoef = data.corr()
CorField = []
for i in CorrKoef:
    for j in CorrKoef.index[CorrKoef[i] > 0.95]:
        if i <> j and j not in CorField and i not in CorField:
            CorField.append(j)
            print "%s-->%s: r^2=%f" % (i,j, CorrKoef[i][CorrKoef.index==j].values[0])
```

```
Electr-->UdN: r^2=0.957896
```

Рисунок 2 - Алгоритм для выявления коррелирующих предикторов

По результатам проверки были выявлены предикторы, имеющие корреляцию более 95% с другими параметрами ('UdCmod', 'AllDolom', 'CaO121', 'UdGaz'). Данные предикторы были исключены из построения модели для предотвращения переобучения. Таким образом, в модель вошли 36 предикторов (рис. 3).

```
sel_vars=['Time', 'Electr', 'Nsr', 'Wslit', 'UdN', 'Badya', 'Lom4SH', 'Lom3SH', 'Lom15', 'Chugun',
          'Lom905', 'Lom2SH', 'LomHBI', 'AllLom', 'Exit', 'O2mod', 'O2modXX', 'UdO2', 'GazMod', 'Cmod',
          'CdspRMH', 'AllC', 'Allizv', 'Udizv', 'Tvipusk', 'Trasch', 'KonO2', 'C111',
          'FeSiMn', 'Al160', 'FeSi65', 'DoloSiroj', 'BKKP4', 'Olive', 'Boloto', 'VesShlak']
```

Рисунок 3 - Предикторы для построения модели

Для построения и проверки качества работы модели основная выборка была разделена в соответствии с требованиями для построения моделей машинного обучения на тренировочную и тестовую выборки. Валидационная выборка отдельно в данной задаче не выделялась в связи с крайне ограниченным набором данных для построения модели. Особенностью сбора информации для основного набора данных являлась последовательная запись о взятии проб от плавки к плавке, поэтому крайне важно было использовать для разделения выборок не кросс-валидацию, а выделение некоторой доли экспериментальных данных с «хвоста» датафрейма. Для проверки качества работы модели взяли 25% от основного набора данных. Таким образом, после разделения выборок получили 4 массива для построения и тестирования работы модели (рис.4)

 ISSN 1562-9945

```
train_data = data.iloc[:-int(0.25*data.shape[0]), :]
test_data = data.iloc[-int(0.25*data.shape[0]):, :]
train_labels = targets.iloc[:-int(0.25*data.shape[0])]
test_labels = targets.iloc[-int(0.25*data.shape[0]):]
```

```
print data.shape, train_data.shape, test_data.shape, train_labels.shape, test_labels.shape
```

```
(105, 36) (78, 36) (27, 36) (78L,) (27L,)
```

Рисунок 4 - Разделение выборки на train и test

Первоначально для моделирования был выбран метод логистической регрессии, который хорошо работает с разреженными данными и небольшими выборками. Однако, даже при использовании L2-регуляризации и нивелировании балансировки классов, модель, построенная при помощи данного метода моментально переобучилась и была фактически бесполезна на тестовых данных. Полученные ROC-кривые и коэффициенты Gini для тренировочной и тестовой выборок представлены на рис. 5.

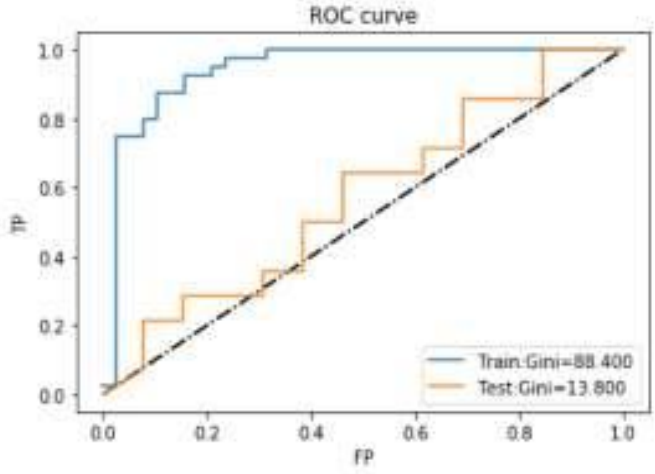


Рисунок 5 - ROC-кривые и коэффициенты Gini для тренировочной (Train - голубая линия) и тестовой (Test – оранжевая линия) выборок, полученные по модели логистической регрессии. Линия точка-пунктир обозначена ROC-кривая для прогнозирования методом случайного выбора

Показатели Accuracy для модели также оказались соответствующими — для тренировочной выборки 0.859 и 0.519 для тестовой, что близко к случайному выбору.

Следует отметить, что модели логистической регрессии часто подразумевают, что предикторы имеют некоторую взаимосвязь, в отличие от моделей, основанных, например, на деревьях решений. Поэтому следующим методом, который был использован для построения

модели выбрали классификатор на основе случайного леса, для которого было подобрано оптимальное количество деревьев (рис. 6).

```
rf_classifier = ensemble.RandomForestClassifier(n_estimators = 20)
rf_classifier.fit(train_data, train_labels)
```

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=20, n_jobs=1,
            oob_score=False, random_state=None, verbose=0,
            warm_start=False)
```

Рисунок 6 - Параметры классификатора случайного леса

Данный метод оказался более подходящим для классификации исходных данных — точность классификатора составила 0.926 на тестовых данных. На первом месте среди топа предикторов с максимальным весом после сортировки оказался предиктор Time (время нахождения расплава в печи) с весом около 0.59 (рис. 7).

```
ii=[]
jj=[]
for i,j in enumerate(rf_classifier.feature_importances_):
    ii.append(data.columns[i])
    jj.append(j)
p=pd.DataFrame(data={'Numb': ii, 'Weigth': jj})
print(p.sort_values(by='Weigth', ascending=False)[:10])
```

```
        Numb    Weigth
0       Time  0.590471
35  VesShlak  0.024976
34    Boloto  0.022777
14      Exit  0.021375
7     Lom3SH  0.020634
16   O2modXX  0.020147
2        Nsr  0.019770
24   Tvipusk  0.018904
20   CdspRMH  0.018883
3      Wslit  0.017550
```

Рисунок 7 - Топ предикторов с весами в модели случайного леса

Абсолютная предикативность данного параметра полностью соответствует представлениям о важности времени при протекании физико-химических реакций во время выплавки стали. Следом за временем расположились предикторы, характеризующие количество подаваемых и расплавляемых в печи материалов. Приведенные веса параметров модели указывают на их высокую предикативность для по-

лучения прогнозируемого класса, исходя из требований, предъявляемых к моделям машинного обучения [3].

Полученные ROC-кривые и коэффициенты Gini для тренировочной и тестовой выборок по новой модели представлены на рис. 8.

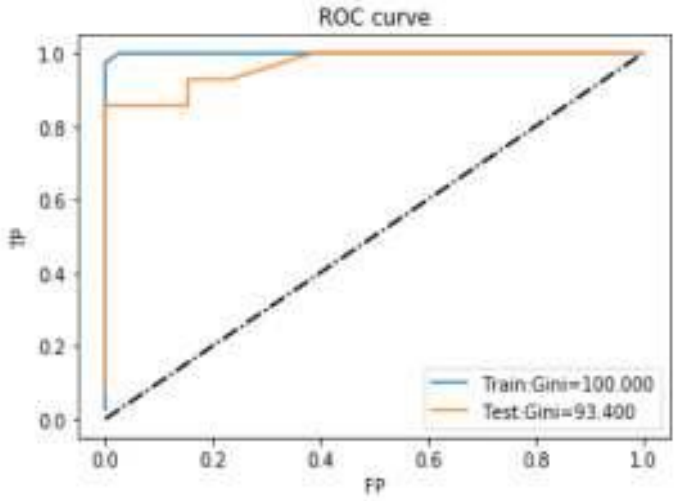


Рисунок 8 - ROC-кривые и коэффициенты Gini для тренировочной и тестовой выборок, полученные по модели случайного леса

На тренировочной выборке отмечается некоторое переобучение модели, что объясняется недостаточным объемом тренировочной выборки, которая была предоставлена для построения модели. В дальнейшем с увеличением датасета и повторной настройкой параметров модели, данная проблема будет устранена. Вместе с этим модель показывает хорошие результаты при тестировании, а разница между коэффициентами Gini не превышает 7%, что является допустимым.

Для моделей бинарной классификации чаще всего характерно применение атрибута predict_proba, который предсказывает не класс, а вероятность отнесения к одному из классов, нежели атрибута predict, предсказывающего номер класса. Так предсказанные вероятности по всем объектам из выборки ранжируют в соответствии с предсказанными вероятностями и получают равные (по количеству объектов) интервалы (децили). Результаты такого ранжирования для тренировочной выборки на 8 интервалов приведены на рис. 9.

| BG decil | 0 | 1 | prob_min | prob_max | sum | Good % | Bad % |
|---|---|---|---|---|---|---|---|
| 0.0 | 11 | 0 | 0.05 | 0.10 | 11 | 0.00 | 100.00 |
| 1.0 | 15 | 0 | 0.15 | 0.20 | 15 | 0.00 | 100.00 |
| 2.0 | 4 | 0 | 0.25 | 0.25 | 4 | 0.00 | 100.00 |
| 3.0 | 6 | 1 | 0.30 | 0.50 | 7 | 14.29 | 85.71 |
| 4.0 | 0 | 11 | 0.55 | 0.70 | 11 | 100.00 | 0.00 |
| 5.0 | 0 | 10 | 0.75 | 0.85 | 10 | 100.00 | 0.00 |
| 6.0 | 0 | 17 | 0.90 | 0.95 | 17 | 100.00 | 0.00 |
| 7.0 | 0 | 1 | 1.00 | 1.00 | 1 | 100.00 | 0.00 |

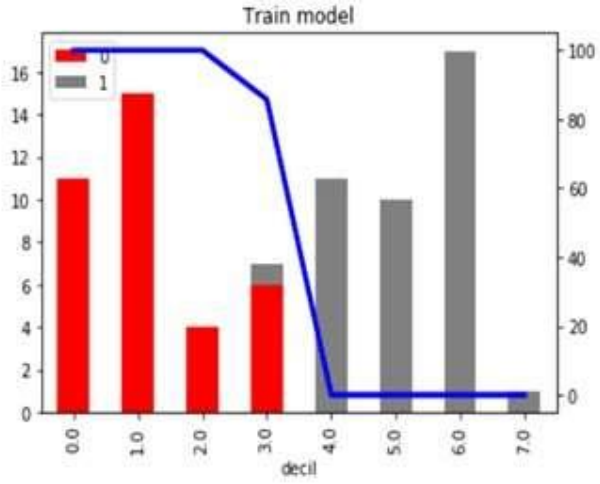


Рисунок 9 - Результаты ранжирования тренировочной выборки по вероятностям отнесения к классу 1 (Good)

В результате такого ранжирования для каждого интервала вероятностей определяется дальнейшие действия в соответствии с технологическим процессом производства или бизнес-задачей. В нашем случае, исходя из полученных распределений вероятности:

– если в результате моделирования получена вероятность <= 0.25 (попадание в 0-2 интервалы вероятностей), то расплав на данный момент не готов к выпуску по химическому составу, следует повторить пересчет вероятности позже;

– если в результате моделирования получена вероятность >= 0.55 (попадание в 4-7 интервалы вероятностей), то расплав готов к выпуску по химическому составу, можно останавливать плавку или делать дополнительные уточняющие анализы химического состава в случае необходимости;

– если в результате моделирования получена вероятность > 0.25 и < 0.55 (попадание в 3 интервал вероятностей), то с вероятностью более 0.85 данный расплав не готов по химическому составу и рекомендуется в скором времени повторить пересчет вероятности.

Работа модели была проверена на тестовой выборке, на которой модель не обучалась. В результате получили следующую таблицу и график распределения вероятностей (рис. 10).

 ISSN 1562-9945

| BG<br>decil | 0 | 1 | score_min | score_max | sum | Good % | Bad % |
|---|---|---|---|---|---|---|---|
| 1 | 6 | 0 | 0.10 | 0.20 | 6 | 0.00 | 100.00 |
| 2 | 2 | 0 | 0.25 | 0.25 | 2 | 0.00 | 100.00 |
| 3 | 5 | 2 | 0.30 | 0.50 | 7 | 28.57 | 71.43 |
| 4 | 0 | 5 | 0.60 | 0.70 | 5 | 100.00 | 0.00 |
| 5 | 0 | 6 | 0.75 | 0.85 | 6 | 100.00 | 0.00 |
| 6 | 0 | 1 | 0.95 | 0.95 | 1 | 100.00 | 0.00 |

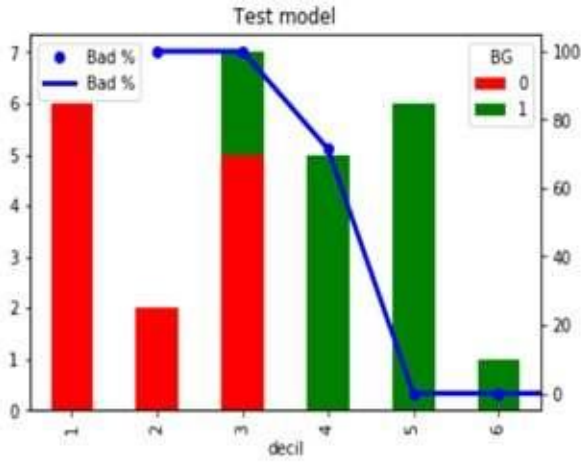


Рисунок 10 - Результаты ранжирования тестовой выборки по вероятностям отнесения к классу 1 (Good)

По результатом подсчета вероятностей для объектов из тестовой выборки оказалось, что у нас отсутствуют объекты, вероятность по которым попадает в 0-й и 7-й интервалы, однако, по остальным интервалам соотношения объектов разных классов соответствуют тренировочной выборке. Небольшие отличия в граничных вероятностях каждого интервала выровняются после добавления большего количества объектов для построения модели. Полученные по тренировочной (рис. 11) и тестовой (рис. 12) выборкам метрики качества также подтверждают адекватность и работоспособность модели уже на данном этапе.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Good | 0.97 | 1.00 | 0.99 | 38 |
| Bad | 1.00 | 0.97 | 0.99 | 40 |
| avg / total | 0.99 | 0.99 | 0.99 | 78 |

Рисунок 11 - Метрики качества для тренировочной выборки

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Good | 0.87 | 1.00 | 0.93 | 13 |
| Bad | 1.00 | 0.86 | 0.92 | 14 |
| avg / total | 0.94 | 0.93 | 0.93 | 27 |

Рисунок 12 - Метрики качества для тестовой выборки

Выполненные моделирование и тестирование модели подтвердили актуальность и перспективность использования методов machine learning для прогнозирования этапов или результатов технологических процессов, в том числе в металлургии. Внедрение данной модели

на производстве позволит ограничить количество проб для определения химического состава расплава до одного раза за плавку, что сэкономит материальные затраты на проведение анализа в лабораториях, в целом сократит время проведения плавки, а также исключить влияние человеческого фактора на результат плавки.

Дальнейшее развитие решения рассмотренной в статье задачи видится в дополнении исходных данных для построения модели, проведении более масштабных экспериментальных исследований и добавлении новых предикторов в модель. Также возможно переформулирование данной задачи классификации в задачу регрессии для вычисления по параметрам загружаемого материала и техническим характеристикам оборудования времени нахождения расплава в печи для получения требуемого химического состава. При реализации модели прогнозирования оставшегося времени «доводки» расплава также рекомендуется автоматический пересчет результата моделирования после изменения любого из предикторов модели (например, после добавления новых порций материалов в печь). Одновременно, после формирования достаточного объема исходных данных, рекомендуется изменение инструмента для построения модели на более прогрессивные библиотеки, работающие по принципу деревьев решений, например, XGBoost [4] или LightGBM [5], которые при правильной настройке не будут переобучаться на тренировочных данных и доказали стабильность своей работы в других бизнес-задачах.

## ЛИТЕРАТУРА

1. Создание и внедрение инновационных технологий электросталеплавильного производства легированных сталей специального назначения, 2018. URL: http://www.golos.com.ua/rus/article/307447. (Дата обращения: 08.10.2018).
2. Машинное обучение, 2018. URL: https://ru.wikipedia.org/wiki/Машинное_обучение (Дата обращения: 17.10.2018).
3. Feature Selection Using Random Forest, 2017. URL: https://chrisalbon.com/machine_learning/trees_and_forests/feature_selection_using_random_forest/ (Дата обращения: 09.10.2018).
4. XGBoost Documentation, 2016. URL: https://xgboost.readthedocs.io/en/latest/ (Дата обращения: 17.10.2018).
5. Welcome to LightGBM's documentation!, 2018. URL: https://lightgbm.readthedocs.io/en/latest/ (Дата обращения: 17.10.2018).