М.С. Ларькин, Н.М. Гаврилова, И.А. Донкова

Тюменский государственный университет, г. Тюмень

УДК 004.048

ПОСТРОЕНИЕ РЕГРЕССИОННОЙ МОДЕЛИ ДЛЯ ПРОГНОЗИРОВАНИЯ ПЛАНОВОГО СРОКА ВЫПОЛНЕНИЯ ЗАДАЧ В DIRECTUM RX

Аннотация. Основой системы расчета планового срока выполнения задач является прогностическая модель, позволяющая определить необходимое количество рабочих часов, которые понадобятся для выполнения задачи. В данной статье проведено сравнение различных прогностических моделей и проведена оценка качества рассматриваемых моделей.

Ключевые слова: DIRECTUM RX, случайный лес, машинное обучение, линейная регрессия, полиномиальная регрессия, регрессия со штрафом.

Введение

Современные системы электронного документооборота (СЭД) для своей работы часто используют методы машинного обучения, однако в основном с помощью этих средств решаются такие задачи, как классификация
документов [1].

На основе такого подхода предлагается использовать методы машинного обучения в качестве способа для расчета планового срока выполнения задачи.

Для разработки системы расчета планового срока выполнения задач, необходимо было построить модель, которая будет прогнозировать количество рабочих часов, необходимых для выполнения поставленной задачи на основе некоторого набора факторов.

Данная задача является задачей регрессии, которая состоит в прогнозировании количественного признака объекта на основании прочих его признаков [2].

Подходы к оцениванию многофакторных моделей

При оценке многофакторных моделей необходимо проверить как связаны между собой объясняющие переменные.

Корреляционный анализ выполнен на основе вычислений матриц парных коэффициентов корреляции и детерминации.

Коэффициент корреляции Пирсона (r) рассчитывается по формуле [3]:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}}, \quad (1)$$

где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \;\; \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Коэффициент детерминации (r^2) представляет собой долю дисперсии y, которая обусловлена линейным соотношением с x.

Для многофакторных моделей применяют матрицы, составленные из коэффициентов парных корреляций (рисунок 1) [4].

$$R = \begin{array}{c} \\ y \\ x_1 \\ - \\ x_m \end{array} \begin{array}{c} \begin{array}{ccccc} y & x_1 & - & x_m \end{array} \\ \begin{pmatrix} 1 & r_{yx_1} & r_{y_2x_2} & \cdots & r_{y_mx_m} \\ r_{x_1y} & 1 & r_{x_1x_2} & \cdots & r_{x_1x_m} \\ - & - & - & - & - \\ r_{x_my} & r_{x_mx_1} & r_{x_nx_2} & \cdots & 1 \end{pmatrix} \end{array}$$

Рис.1. Матрица коэффициентов корреляции

Для оценки модели в целом наиболее популярными метриками являются следующие метрики.

Совокупный коэффициент детерминации (R^2) – данная метрика показывает какая доля изменения исследуемого признака учтена в модели. Совокупный коэффициент детерминации рассчитывается по формуле [5]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad \text{где} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad (3)$$

Средняя абсолютная ошибка (MAE) – данная метрика измеряет среднюю сумму абсолютной разницы между фактическими значениями и прогнозируемыми значениями. Чем меньше данный показатель, тем лучше модель прогнозирует значение целевой переменной. Средняя абсолютная ошибка рассчитывается по формуле [6,7]:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}| \quad (4)$$

Средняя абсолютная ошибка в процентном соотношении (MAE%) – данная метрика позволяет оценить в процентном соотношении среднюю абсолютную ошибку, относительно среднего значения фактического значения [6,7]:

$$\frac{\sum_{i=1}^{n}|y_i - \hat{y}|}{\sum_{i=1}^{n}|y_i|} * 100\% \quad (5)$$

Среднеквадратическая ошибка (RMSE) – данная метрика показывает среднюю сумму разностей между фактическими значениями и прогнозируемыми. Среднеквадратическая ошибка рассчитывается по формуле [6,7]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2} \quad (6)$$

Среднеквадратическая ошибка (RMSE) в отличие от MAE менее устойчива к выбросам, а за счет возведения разницы между фактическим и прогнозируемым значением в квадрат, придает больший вес большим ошибкам.

Отбор факторов регрессионной модели

Исходные данные были получены из СЭД DIRECTUM RX и содержали в себе информацию об инициаторе задачи, исполнителе, вложениях в задачу и виде работ.

В качестве результирующего признака (y) использовалось количество рабочих часов, затраченных на выполнение задачи с точностью до 1 знака после запятой.

Для устранения явления мультиколлинеарности была проанализирована матрица выборочных парных коэффициентов корреляции (рисунок 1), построенная с помощью инструментов языка программирования Python.

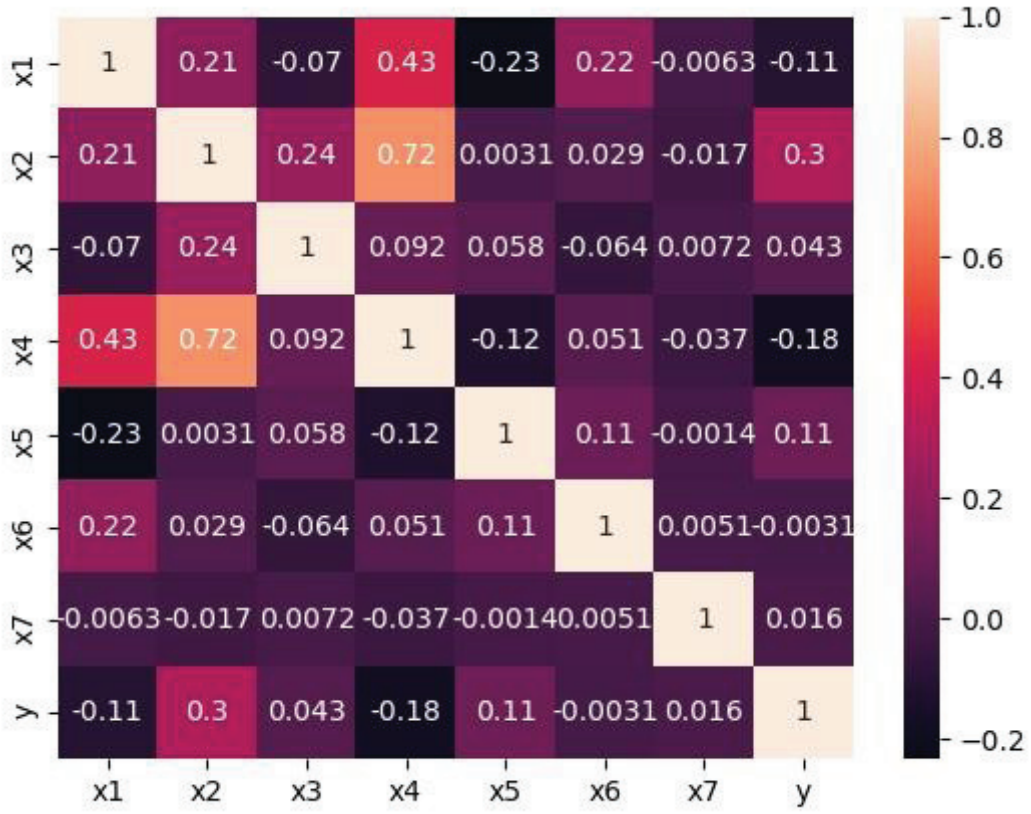

Рис. 1. Матрица корреляции.

На основе полученных значений коэффициентов парной корреляции, можно сделать вывод, что фактор x6 сильно коррелирует с фактором x2, при этом фактор x2 сильно коррелирует с результирующим признаком y. Поэтому признак x6 можно исключить из исходного набора данных.

Таким образом, после предобработки исходных данных был получен набор данных, состоящий 8171 примера с 30 факторными признаками и одной целевой переменной, который можно использовать для построения моделей прогнозирования срока выполнения задачи.

В качестве инструмента для работы с данными использовался язык программирования Python и его библиотеки «pandas» и «scikit-learn».

Сравнение регрессионных моделей

После предобработки данные были разделены на обучающую и тестовую выборку в процентном соотношении 80% к 20%. Таким образом, обучающая выборка состояла из 6536 примеров, а тестовая состояла из 1635 примеров. На основе обучающей выборки были построены такие модели:

- линейная регрессионная модель;
- полиномиальная модель (со степенью 2 и 3);
- полиномиальная модель (со степенью 2 и 3) с регуляризацией (LASSO и Ridge);
- модель, построенная с помощью метода случайного леса.

При сравнении моделей использовалась кросс-валидация по K-блокам.

Кросс-валидация – это способ проверки того, насколько успешно применяемый в модели статистический анализ способен работать на независимом наборе данных. Существуют различные типы кросс-валидации [8], такие как кросс-валидация по K-блокам, валидация последовательным случайным сэмплированием и т.д.

K-блочная кросс-валидация представляет собой разбиение набора данных на тренировочный и тестовый наборы. Для каждого такого разбиения модель обучается на тренировочных данных, а точность предсказания оценивается на тестовом наборе. После проведения предсказаний результаты усредняются по всем разбиениям.

Сравнение моделей представлено в таблице 1

Условные обозначения для таблиц 1 и 2:

- LR – линейная регрессионная модель;
- PR (2) – полиномиальная модель степени 2;
- PR (3) – полиномиальная модель степени 3;

- PR L (2) – полиномиальная модель степени 2, построенная с помощью метода LASSO регрессии;
- PR L (3) – полиномиальная модель степени 3 построенная с помощью метода LASSO регрессии;
- PR R (2) – полиномиальная модель степени 2 построенная с помощью метода Ridge регрессии;
- PR R (3) – полиномиальная модель степени 3 построенная с помощью метода Ridge регрессии;
- RF (200) – модель, построенная с помощью метода случайного леса, состоящим из 200 деревьев;
- с.к.ч. – среднее количество часов.

Таблица 1. Сравнение качества моделей

	LR	PR (2)	PR (3)	PR L (2)	PR L (3)	PR R (2)	PR R (3)	RF (200)
R^2	0,56	-3,3E+9	-2,2E+15	0,74	0,79	0,74	0,64	0,86
MAE (часы)	10,5	873,5	3E+7	6,74	6,00	6,74	6,53	4,46
MAE %	63 %	5293 %	1,8E+8 %	41%	36 %	41 %	39 %	27 %
RMSE (часы)	14,9	6E+5	1,1E+9	11,5	10,3	11,24	10,33	8,27

На основе полученных данных можно сделать вывод, что из представленных моделей лучше всего данным соответствует модель случайного леса. В то же время, модель полиномиальной регрессии без применения регуляризации не соответствует исходным данным, вследствие чего средняя абсолютная ошибка является самой большой по сравнению с остальными рассматриваемыми моделями.

После отбора лучшей модели на основе кросс-валидации итоговая модель была обучена на полном наборе данных.

Данная модель была построена с помощью метода случайного леса и состояла из 200 деревьев решений.

Оценка качества модели на целом тестовом наборе данных показала следующие результаты:

5. $R^2 = 0{,}87$
6. MAE (часы) = 4,24
7. MAE% = 25%

Помимо получения общих оценок был проведен анализ допустимости полученной ошибки по конкретному виду работ (таблица 2).

Таблица 2. Ошибка модели в зависимости от вида работы

	Разработка	Внутреннее взаимодействие	Сопровождение
MAE (часы)	4,0	3,74	4,7
MAE %	26 %	23,9 %	24 %
с.к.ч.	15,4	15,8	19,8

Так, например, для задач с видом работ «Разработка» среднее время выполнения составляет 15,4 часа, с видом работ «Внутреннее взаимодействие» - 15,8, а для задач с типом работ «Сопровождение» - 19,8.

Таким образом, для типа работ «Разработка» средняя ошибка составляет 26%, для вида работ «Внутреннее взаимодействие» - 23,9%, а для вида работ «Сопровождение» - 26%.

На основе полученных данных можно сделать вывод о том, что средняя ошибка в зависимости от вида работ не превышает 27%, а в среднем составляет 25%. Это является допустимым, так как в процессе работы над задачей у исполнителя меняется загруженность, что невозможно оценить на этапе создания задачи.

Заключение

В статье рассмотрено построение и сравнение таких регрессионных моделей, как линейная регрессионная модель, полиномиальная модель (со степенью 2 и 3), полиномиальная модель (со степенью 2 и 3) с регуляризацией (LASSO и Ridge) и модель, построенная с помощью метода случайного леса. По результатам сравнения на основе корреляционного анализа в качестве прогнозной модели выбрана модель случайного леса, которая будет использована при прогнозировании планового срока выполнения задач в DIRECTUM RX.

СПИСОК ЛИТЕРАТУРЫ

1. Мандрыгина В.С. Возможности применения технологий искусственного интеллекта в задачах автоматизации электронного документооборота предприятий на примере WSSDocs. Екатеринбург: Уральский государственный экономический университет, 2019. 3 с.
2. Открытый курс машинного обучения [Электронный ресурс]: – Электрон. текстовые дан. – Режим доступа: https://habr.com/ru/company/ods/blog/322534/. свободный. – Загл. с экрана (дата обращения: 06.06.2020).
3. Коэффициент корреляции Пирсона [Электронный ресурс]: – Электрон. текстовые дан. – Режим доступа: http://cito-web.yspu.org/link1/metod/met125/node35.html. свободный. – Загл. с экрана (дата обращения: 06.06.2020).
4. Анализ матрицы парных коэффициентов корреляции [Электронный ресурс]: – Электрон. текстовые дан. – Режим доступа: https://studopedia.info/4-5062.html. свободный. – Загл. с экрана (дата обращения: 06.06.2020).
5. Оценка результатов линейной регрессии. [Электронный ресурс]: – Электрон. текстовые дан. – Режим доступа:

https://habr.com/ru/post/195146/. свободный. – Загл. с экрана (дата обращения: 06.06.2020).

6. Основные оценки точности прогнозирования временных рядов. [Электронный ресурс]: – Электрон. текстовые дан. – Режим доступа: https://www.mbureau.ru/blog/osnovnye-ocenki-tochnosti-prognozirovaniya-vremennyh-ryadov. свободный. – Загл. с экрана (дата обращения: 06.06.2020).
7. MAPE v/s MAE% v/s RMSE. [Электронный ресурс]: – Электрон. текстовые дан. – Режим доступа: https://medium.com/@agrimabahl/mape-v-s-mae-v-s-rmse-3e358fd58f65. свободный. – Загл. с экрана (дата обращения: 06.06.2020).
8. Сравнение различных видов кросс-валидации. [Электронный ресурс]: – Электрон. текстовые дан. – Режим доступа: http://datareview.info/article/sravnenie-razlichnyih-vidov-kross-validatsii/. свободный. – Загл. с экрана (дата обращения: 06.06.2020).