В.В. Паседько; А.И. Якимов, д.т.н., доц.; Н.В. Выговская
(Белорусско-Российский университет, г. Могилев)

РАЗРАБОТКА МОДЕЛИ НА ОСНОВЕ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ ПРИ АНАЛИЗЕ МЕДИЦИНСКИХ ДАННЫХ

Логистическая регрессия или логит регрессия (англ. logit model) – это статистическая модель, используемая для предсказания вероятности возникновения некоторого события путём подгонки данных к логистической кривой. Метод, основанный на применении логистической регрессии, является одним из самых используемых при решении проблемы классификации [1].

Была поставлена задача проанализировать медицинские данные пациенток с опухолями в грудной массе, которые были собраны в штате Висконсин, США. Данные были взяты из Kaggle (https://www.kaggle.com) – системы организации конкурсов по исследованию данных, а также социальной сети специалистов по обработке данных и машинному обучению. Данные были получены из оцифрованного изображения биопсии грудной массы, также они были собраны доктором Уильямом Х. Вольбергом [2] в университете Висконсин, больница Мэдисон, США. Они представляют собой характеристики ядер клеток.

Для анализа данных использовался язык программирования Python и его библиотеки для визуализации. В качестве метода для построения предиктивной математической модели был выбран метод логистической регрессии, который также был имплементирован на языке Python [3]. Целью разработки программного обеспечения (ПО) было построение модели, способной прогнозировать вероятность рака груди на новых данных. Для прогнозирования требуется взять биопсию и оцифровать её изображение. По этим новым данным можно будет прогнозировать вероятность рака груди у пациента.

В процессе работы был проведён разведывательный анализ данных, а именно:

- рассчитаны статистические показатели для каждой переменной-предиктора;

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000
mean	0.372583	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	0.048919	0.181162
std	0.483918	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0.038803	0.027414
min	0.000000	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.000000	0.106000
25%	0.000000	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0.020310	0.161900
50%	0.000000	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	0.033500	0.179200
75%	1.000000	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.074000	0.195700
max	1.000000	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201200	0.304000

Рис. 1. Статистические показатели по данным

- рассчитано распределение целевой переменной;

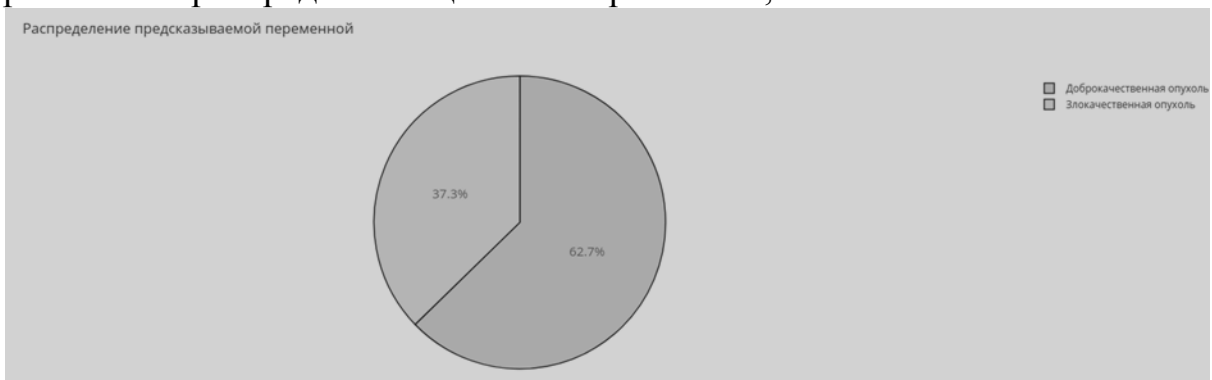

Рис. 2. Распределение целевой переменной

- построены графики распределений переменных-предикторов;

– рассчитана ядерная оценка плотности для переменных-предикторов;

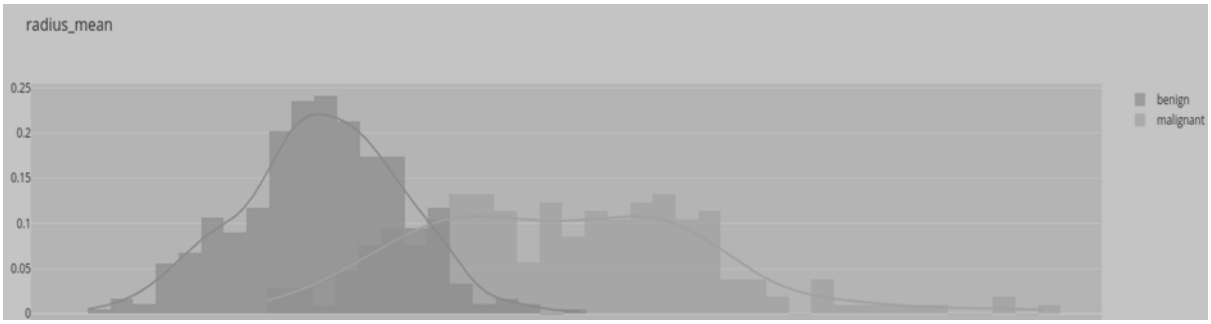

Рис. 3. Распределение и ядерная оценка плотности переменной Radius Mean

– построены графики корреляций между переменными-предикторами, а также между предикторами и целевой переменной;

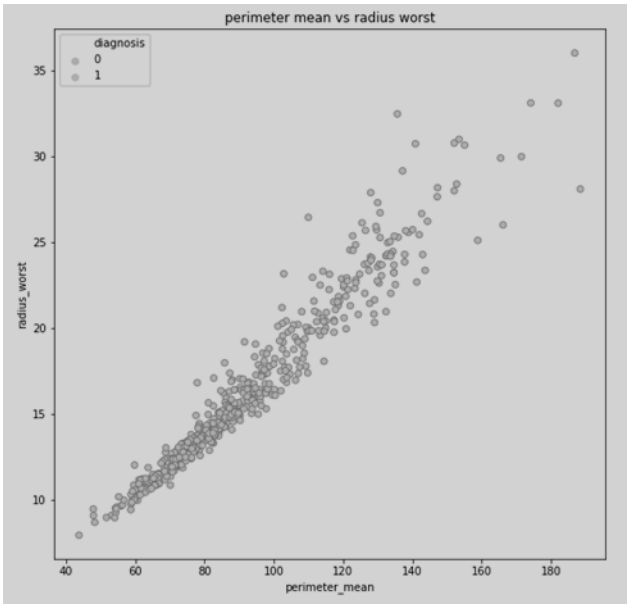

Рис. 4. Корреляция переменных perimeter mean и radius worst

– построен график тепловой карты, который отображает корреляции между переменными;

– найдены позитивно и негативно коррелирующие между собой предикторы.

Далее был выполнен этап подготовки к построению модели, а именно:

– был определен метод логистической регрессии как оптимальный для выполнения задачи классификации;

– был определен метод оценки результата выполнения модели.

Следующим шаг это подготовка датасета:

– была определена матрица переменных-предикторов, а также вектор целевой переменной;

– данные были стандартизированы (Feature Scaling);

– датасет был разделён на тестовую и обучающую выборки;

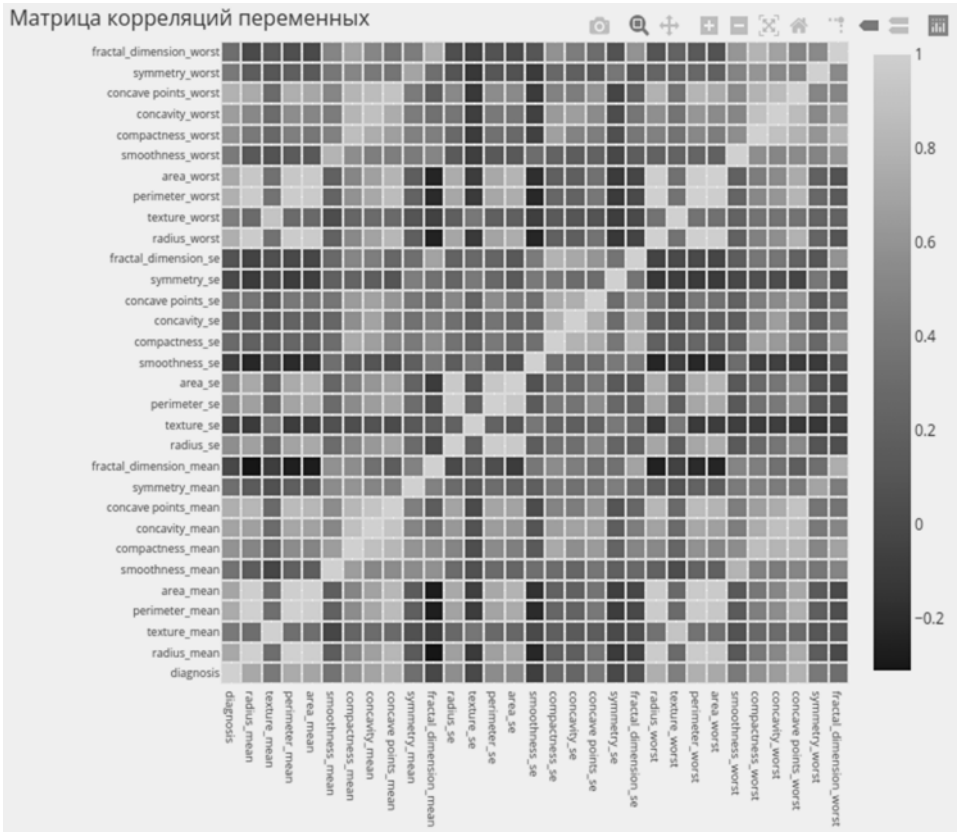

Рис. 5. Матрица корреляций Пирсона для переменных

– были подобраны оптимальные гипер-параметры для построения модели;

– было произведено обучение модели на обучающей выборке, в процессе обучения были подобраны коэффициенты (веса) для каждой переменной-предиктора.

После применения модели на тестовой выборке были получены следующие результаты:

– точность модели на тестовой выборке — 99.4%;

```
Accuracy  =     0.994
Precision =     1.000
Recall    =     0.984
F1_score  =     0.992
```

– точность рассчитывалась по формуле:

$accuracy = (TP + TN)/(TP + TN + FP + FN)$, где TP — количество правильно идентифицированных злокачественных опухолей, TN — количество правильно идентифицированных доброкачественных опухолей, FP — количество ложноотрицательных результатов, FN — количество ложноположительных результатов;

– была построена матрица выполнения модели на тестовой выборке.

– Разработанное ПО может быть использовано при диагностике раковых заболеваний у пациенток с опухолями в грудной массе наряду с другими методами медицинской диагностики. Проведённая работа находится в открытом доступе: https://notebooks.azure.com/viachaslau-pasedzka/projects/data-analysis-for-university

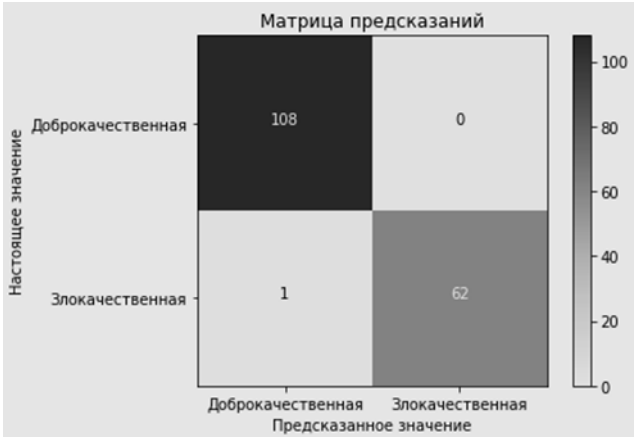

Рис. 6. Матрица выполнения модели на отложенной выборке

Список литературы

1. Леонов, В. Логистическая регрессия в медицине и биологии / В. Леонов [Электрон. ресурс] Биометрика. – 2020. – Режим доступа : http://www.biometrica.tomsk.ru/logit_0.htm – Дата доступа: 10.03.2020.

2. Mangasarian, O. L. Breast Cancer Diagnosis and Prognosis Via Linear Programming / O. L. Mangasarian, W. N. Street, W. H. Wolberg [Электрон. ресурс] Operations research. – 1995. – Vol. 43. – Режим доступа : https://doi.org/10.1287/opre.43.4.578 – Дата доступа: 12.03.2020.

3. Коэльо, Л. П., Вилли Ричарт, В. Построение систем машинного обучения на языке Python / Л. П. Коэльо, В. Ричарт. 2-е изд. / пер. с англ. Слимкина А. А. – М. : ДМК Пресс, 2016. – 302 с. : ил.