UDC 336.774.3

PREDICTING TARGET VARIABLE VALUES AND CONSTRUCTING A ROC -CURVE

D.M. Rahel
PhD in Economics, Associate Professor of the Department of Economics of the Belarusian State University of Informatics and Radioelectronics
ragel@bsuir.by

D.M. Rahel
In 2000 he graduated from the Faculty of Economics of the Belarusian State University of Informatics and Radioelectronics with a degree in Economic Informatics. In 2016 he graduated from the postgraduate course of the Academy of Management under the President of the Republic of Belarus. In 2018 he defended his PhD thesis. Research interests: data mining in marketing, process modeling, data analysis, statistical forecasting, macroeconomics.

Annotation. This article examines approaches to ROC-analysis for modeling commercial data behavior. It examines a set of similar data on retail sales over a specific period of time. Based on existing estimates, a forecast is made and its quality is assessed using baseline metrics confirming the forecast's reliability. Based on the confusion matrix, conclusions were drawn regarding the model's reliability and balance. The final forecast demonstrated a high degree of reliability. The most significant features for classification were identified, providing valuable insights into the factors influencing the final target variable. The combined importance of the identified features allowed us to evaluate their role in the classification process.

Keywords: Data analysis, marketing data, commercial evaluation, big data, data array, forecasting, marketing analytics, commercial analytics, commercial modeling, predictive analytics, behavior forecasting, algorithmic marketing, market forecast, ROC-curve, ROC-analysis.

Introduction. We considered a dataset containing point-of-sale sales results for a major player in the retail market. Based on the initial data, we had to predict the target variable values for the initial dataset and select a forecasting method that would most accurately predict data over the next three months. We also had to evaluate the accuracy and quality of the predictive model. We also had to evaluate the predicted values using a receiver operating characteristic ROC-curve and based on this, identify the three most important predictors influencing the target variable.

Data. In this case, we used commercial sales data with the following parameters:

1. Training set: 10,000 rows, 33 columns.
2. Validation set: 20,000 rows, 33 columns.
3. Target variable: Target (binary: 0 or 1).
4. Distribution of classes in the training set: 50% class 0, 50% class 1.

Method. The Random Forest method was chosen to solve the binary classification problem. This method was chosen for the described dataset for the following reasons:

- Random Forest works effectively with a large number of features and does not require pre-selection of variables;
- the model is resistant to overfitting due to decision tree ensemble;
- the algorithm can account for nonlinear dependencies and interactions between features;

- the method correctly handles data containing gaps and noise, which is especially important for large practical datasets collected at points of sale;
- Random Forest provides a built-in feature importance score, which allows for the interpretation of model results.

Taking this into account, a conclusion was made about the reliability of this method within the framework of the problem being solved.

To assess the quality of the predictive model, a confusion matrix was calculated, along with key classification quality metrics: Accuracy; Precision; Recall; and F1-score. The resulting metric values indicate acceptable quality of the model and its ability to correctly distinguish between objects in the two classes.

Results. The quality of the predictive model was assessed on a validation set representing 20% of the original data, using standard binary classification metrics.

The following performance indicators were obtained:

- Accuracy = 0.6915;
- AUC (Area Under ROC-Curve) = 0.755.

The Accuracy value indicates that the model correctly classifies approximately 69% of observations. An AUC value greater than 0.755 indicates good class discrimination and confirms the adequacy of the constructed model (Figure 1).

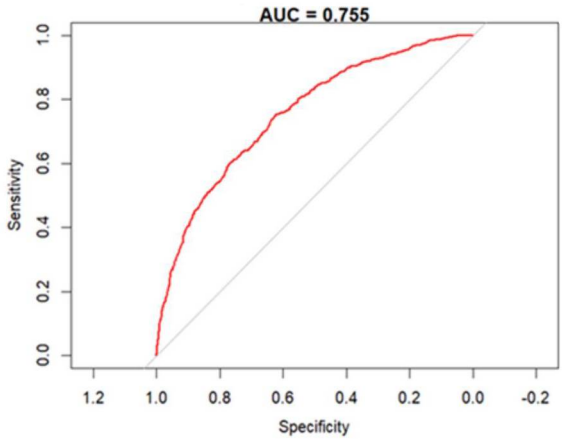

Figure 1. he result of ROC-analysis of the classification of values

Additionally, a confusion matrix was calculated, showing the distribution of correct and incorrect classifications, allowing for a detailed assessment of the model's performance for each class. Based on the feature importance assessment using the Mean Decrease Gini criterion, three of the most significant predictors were identified: P16, P23, and P10. These features contribute most to the decision-making process based on the Random Forest model and have a key influence on the final forecast for our dataset (Figure 2). Taking into account objective assessments of significance, it is these three features that are the most important for forecasting and they will be decisive in forming a forecast for the data set under consideration.

```
Топ-3
> print(head(imp_df, 3))
    Predictor Importance
P16       P16   288.8789
P23       P23   246.7440
P10       P10   226.3062
```

Figure 2. The most significant features in the data set under consideration

In addition, a predictive binary classification model was built using the Random Forest method. The model was trained and tested on a validation set, and its quality was assessed using the Accuracy and AUC metrics. The model was also interpreted by identifying the most important features. The results showed that the model has satisfactory quality and can be used to solve value prediction problems within this dataset (Figure 3).

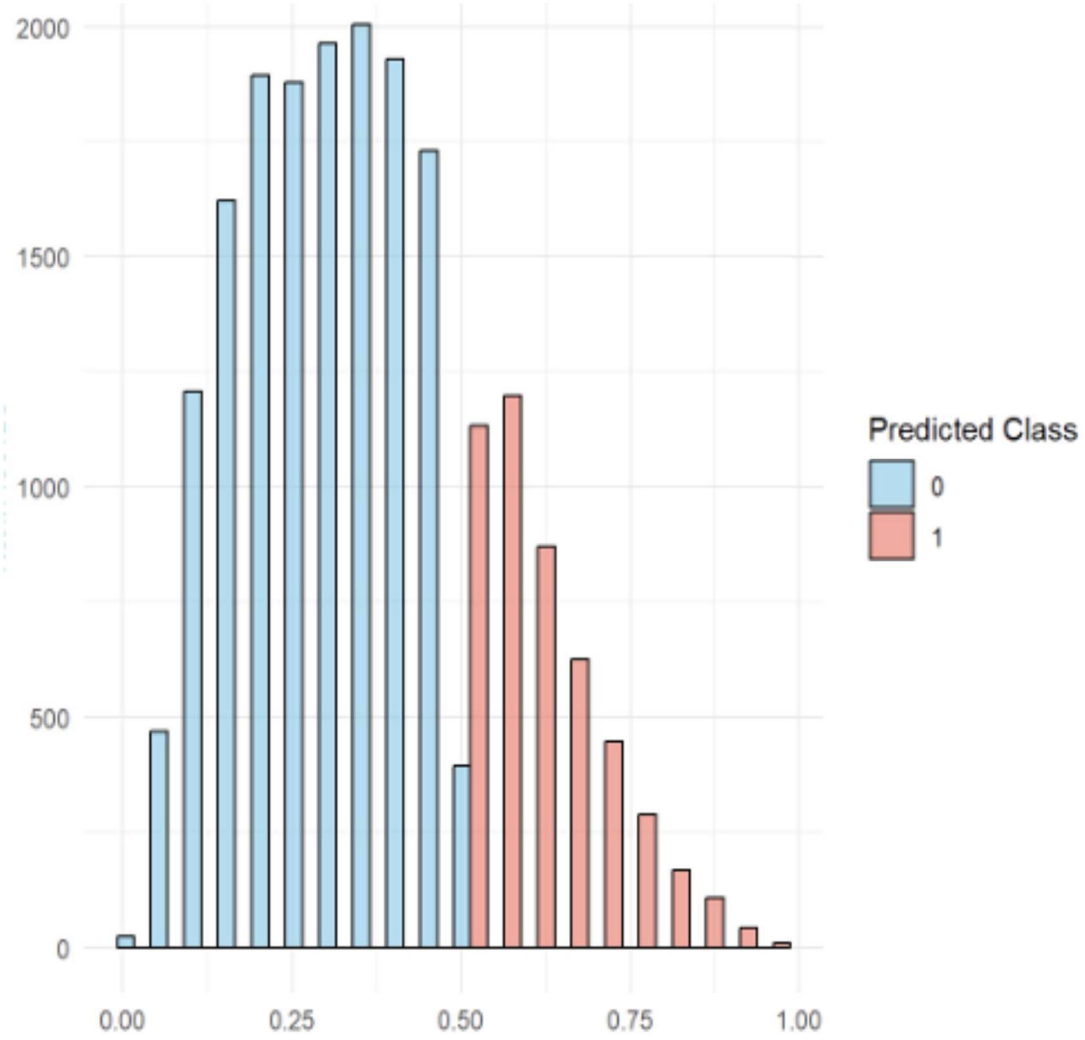

Figure 3. Final class probability distribution

The Random Forest method was chosen to solve the classification problem because it effectively models complex nonlinear relationships, is robust to overfitting, and provides a built-in feature importance assessment mechanism.

The confusion matrix demonstrates the balanced quality of the model: it identifies positive-class objects reasonably well and does not exhibit a critical bias toward false positives.

At the final stage, a final prediction for the dataset was generated, which, following practical implementation, demonstrated a high degree of reliability.

Furthermore, the analysis successfully identified the three most important predictors. Understanding feature importance not only improves the model but also provides valuable insights into the factors influencing the target variable.

The combined importance of the three identified features exceeds 27%, indicating their dominant role in the classification process.

References

[1] Коэн М.И. Прикладная линейная алгебра для исследователей данных / пер. с англ. А.В. Логунова. – М.: ДМК Пресс, 2023. – 328 с.

[2] Практическая статистика для специалистов Data Science: Пер. с англ. / П.Брюс, Э.Брюс. – СПб.: БХВ-Петербург, 2019. – 304 с.

[3] Data Science and Big Data Analytics. A Step by Step Guide to learn Data Science from Scratch with Python Machine Learning and Big Data / Andrew Park. – Published by Andrew Park, 2021. – 124 p.

ПРЕДСКАЗАНИЕ ЗНАЧЕНИЙ ЦЕЛЕВОЙ ПРЕМЕННОЙ И ПОСТРОЕНИЕ ROC-КРИВОЙ

Д.М. Рагель

к.э.н., доцент кафедры экономики Белорусского государственного университета информатики и радиоэлектроники

Аннотация. В статье рассматриваются подходы к ROC-анализу моделирования поведения коммерческих данных. В статье рассматривается набор однотипных данных о продажах в розничной сети в течение отдельно взятого периода времени.

На основании существующих оценок делается прогноз и оценивается его качества на основании базовых показателей, подтверждающих уровень достоверности прогноза. На основании матрицы ошибок сделаны выводы о достоверности и сбалансированности модели.

Итоговый прогноз показал высокую степень достоверности. Определены наиболее значимые признаки для классификации, что позволило получить ценную предметную информацию о факторах, влияющих на итоговую целевую переменную. Совокупная важность выделенных признаков позволила дать оценку их роли в процессе классификации.

Ключевые слова: Анализ данных, маркетинговые данные, коммерческая оценка, большие данные, массив данных, прогнозирование, маркетинговая аналитика, коммерческая аналитика, коммерческое моделирование, прогнозная аналитика, прогнозирование поведения, алгоритмический маркетинг, рыночный прогноз, ROC-кривая, ROC-анализ.