

Лабораторная работа №3. Подбор гиперпараметров модели

ГОРБАН АРТЕМИЙ М8О-307Б-23

ВЫБРАННАЯ МОДЕЛЬ: RANDOM FOREST.

ДАТАСЕТ: ПРОДУКТЫ

Подготовка данных

Цель: Предсказать, популярен ли товар (купили >1000 раз).

Изменения в данных:

Date -> Month, DayOfWeek, DayOfMonth.

Создан таргет is_popular

Признаки: Member_number, Month, DayOfWeek, DayOfMonth.

Метод: Stratified split (80/20).

Гиперпараметры Random Forest

`n_estimators` - Число деревьев

`max_depth` - Макс. глубина дерева

`min_samples_split` - Мин. образцов для разделения узла

`min_samples_leaf` - Мин. образцов в листе

`max_features` - Число признаков для разбиения

`criterion` - Критерий (gini/entropy)

Минусы: Долго при большой сетке.

Методы подбора

Optuna: Интеллектуальный поиск

Плюсы: Эффективно исследует пространство, умный.

Минусы: Сложнее в настройке.

```
def objective(trial):
    """Objective function to minimize"""
    # Hyperparameters
    param1 = trial.suggest_float('param1', 0.1, 1.0)
    param2 = trial.suggest_int('param2', 1, 10)
    param3 = trial.suggest_categorical('param3', ['cat1', 'cat2', 'cat3'])

    # Model training and evaluation
    model = MyModel(param1, param2, param3)
    model.fit(X_train, y_train)
    score = model.score(X_test, y_test)

    # Return the value to be minimized
    return 1 / score

# Optuna optimization
study = optuna.create_study(direction='minimize')
study.optimize(objective, num_trials=100)
```

TRIO (AutoML): Автоматически ищет лучший пайплайн

Плюсы: Полная автоматизация, может найти неочевидное решение.

Минусы: Очень долго, чёрный ящик.

```
from trio import AutoML
from sklearn.datasets import load_iris
from sklearn.metrics import accuracy_score

# Load data
iris = load_iris()
X, y = iris.data, iris.target

# Initialize AutoML
automl = AutoML()

# Fit the model
automl.fit(X, y)

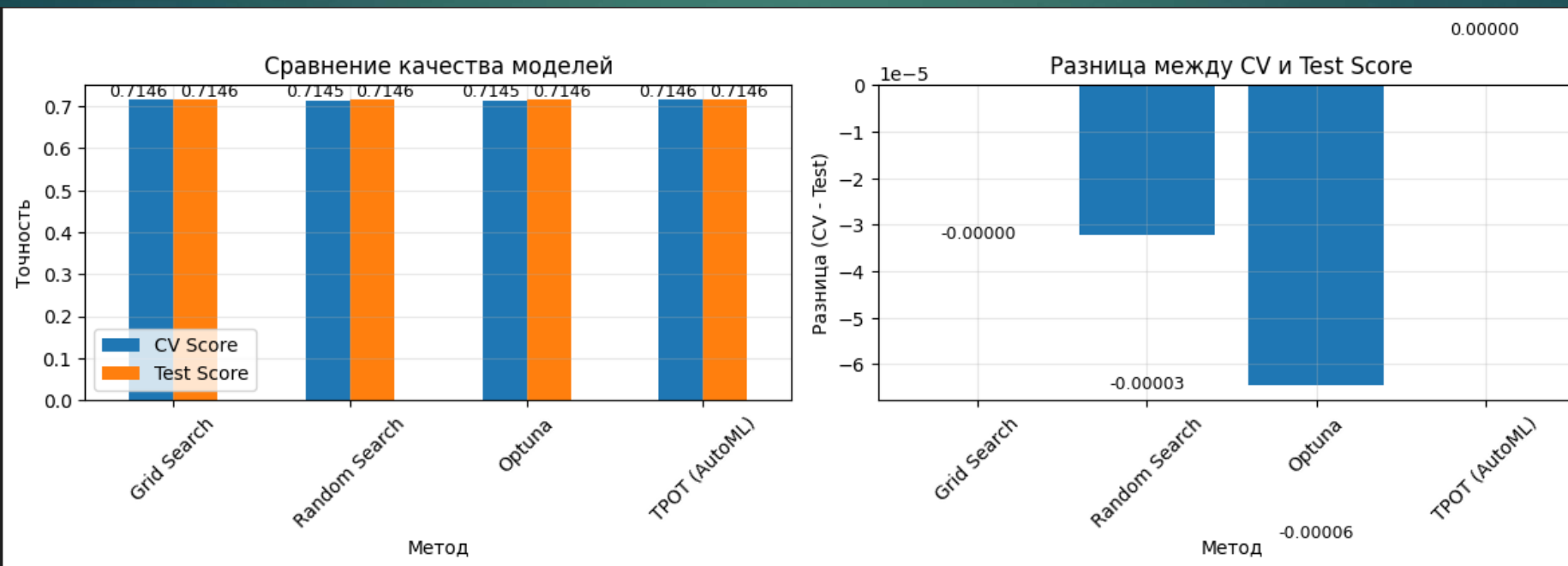
# Get the best model
best_model = automl.get_best_model()

# Evaluate the best model
accuracy = accuracy_score(y, best_model.predict(X))

print(f'Accuracy: {accuracy}')
```

Результаты подбора

	Method	CV Score	Test Score
0	Grid Search	0.714562	0.714562
1	Random Search	0.714530	0.714562
2	Optuna	0.714498	0.714562
3	TPOT (AutoML)	0.714562	0.714562

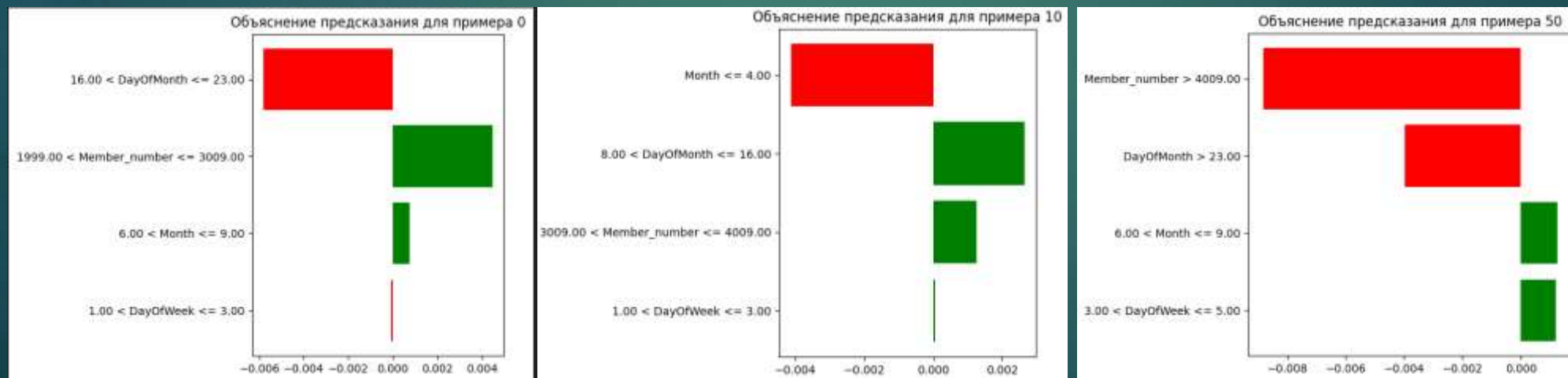


Локальная интерпретация (LIME)

Цель: Объяснить, почему для конкретного клиента модель предсказала популярный или не популярный товар.

LIME создаёт объясняющую модель для конкретного предсказания.

График, показывающий, какие признаки сильнее всего повлияли на это конкретное решение модели

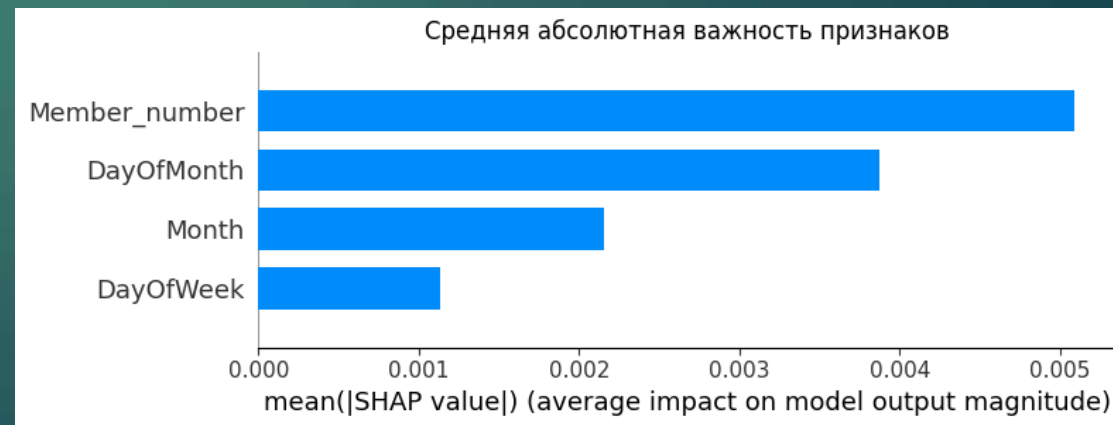
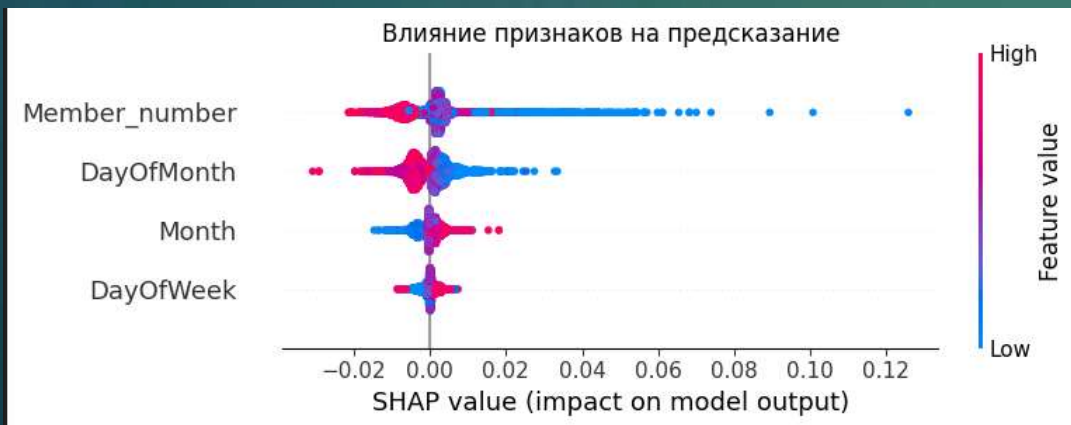


Глобальная интерпретация (SHAP)

Цель: Объяснить, почему для конкретного клиента модель предсказала популярный или не популярный товар.

LIME создаёт объясняющую модель для конкретного предсказания.

График, показывающий, какие признаки сильнее всего повлияли на это конкретное решение модели



ВЫВОДЫ

Grid search оказался лучшим методом подбора гиперпараметров для нашей задачи.

Качество всех методов схожее, но Grid search показала максимальную точность на тесте 0.7146.

Локальная интерпретация (LIME) помогает понять отдельные предсказания модели.

Глобальная интерпретация (SHAP) выявила, что Member_number - ключевой признак для прогноза.

Итог: Работа выполнена, модель обучена, её решения можно объяснить.