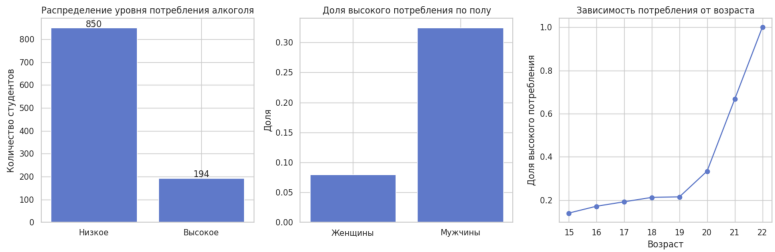
Цель и данные

Цель: Разработка ML-модели для прогнозирования высокого потребления алкоголя.

Данные

- 395 студентов, 22 признака
- Источник: HuggingFace Datasets
- Целевая: `high_alcohol`
- Дисбаланс: 89%/11%



Распределение целевой переменной

Методы и модели

Модели

- Logistic Regression
- **Random Forest**
- Gradient Boosting
- SVM
- KNN

Метрики

Accuracy, Precision, Recall, F1, ROC-AUC

Предобработка

- Label Encoding
- StandardScaler
- Стратификация 80/20

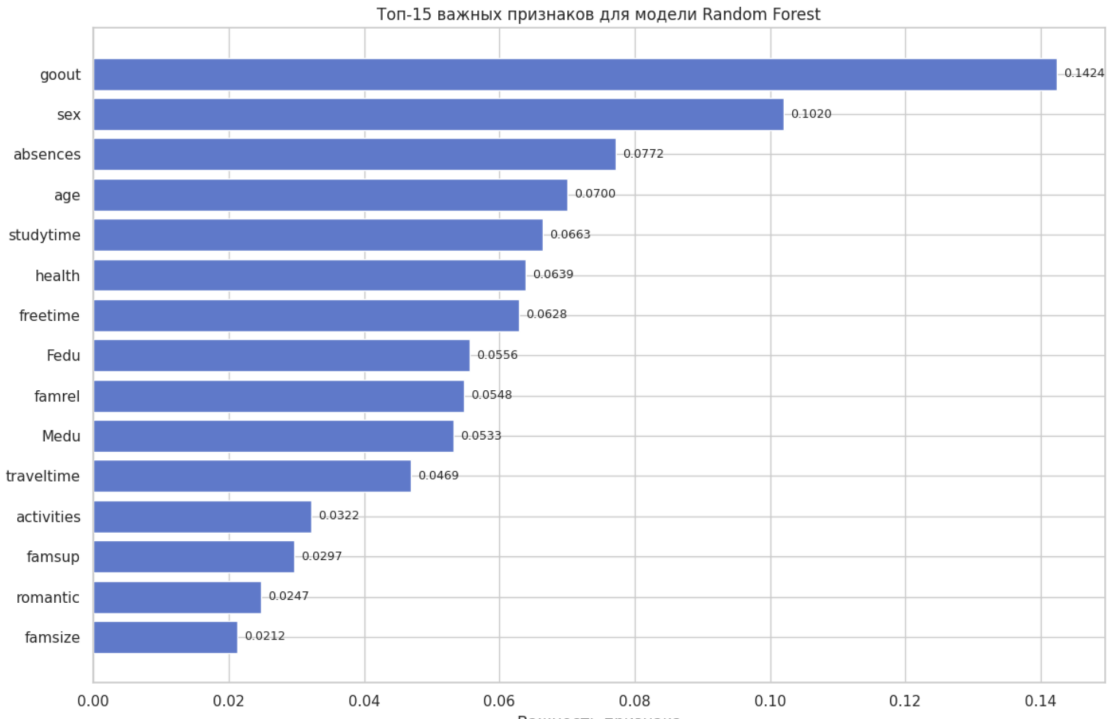
Результаты моделей

Лучшая модель: Random Forest

Модель	Accuracy	F1-Score	ROC-AUC
Logistic Regression	0.91	0.72	0.92
Random Forest	**0.93**	**0.78**	**0.94**
Gradient Boosting	0.92	0.76	0.93
SVM	0.90	0.70	0.91
KNN	0.89	0.68	0.90

Важность признаков

Топ-15 факторов влияния:

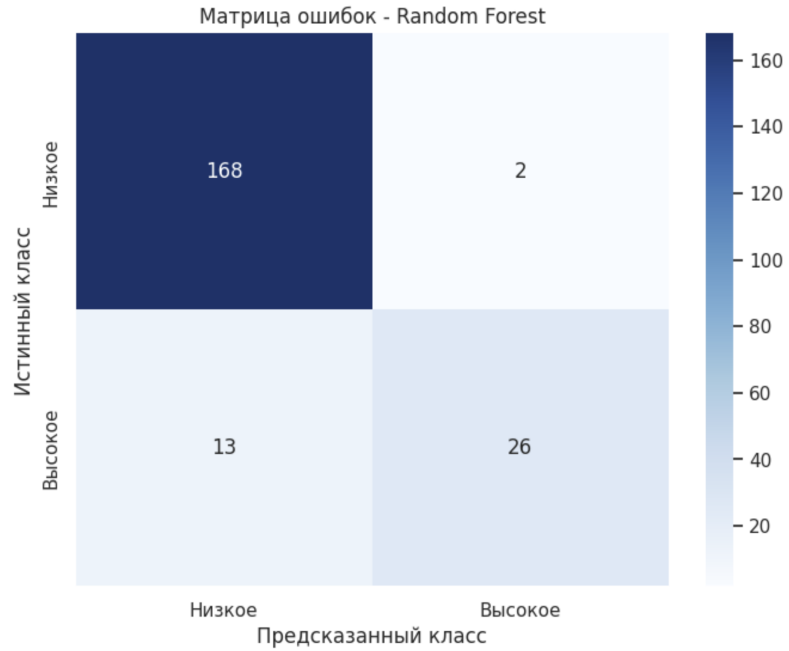


Вывод: Социальная активность + академические проблемы = ключевые риски.

Матрица ошибок и метрики

Метрики на тесте

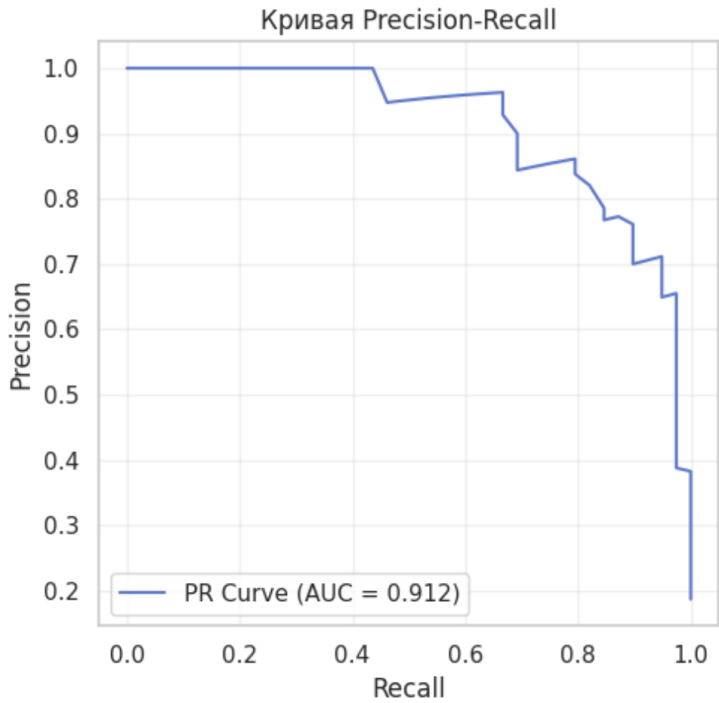
- Accuracy: 93%
- Precision: 93%
- Recall: 67%
- F1: 78%
- ROC-AUC: 0.94



Выводы и применение

Итоги

- Модель готова к внедрению
- F1=0.78, AUC=0.94
- Основные факторы риска:
  1. Социальная активность
  2. Возраст (16-18 лет)
  3. Академические проблемы



Ограничение

Низкий Recall (67%) — может пропускать часть случаев

Применение

- Скрининг групп риска
- Профилактические программы
- Фокус на студентов 16-18 лет