Федеральное государственное автономное

образовательное учреждение высшего образования

«Пермский государственный национальный исследовательский университет» (ПГНИУ)

Региональный институт непрерывного образования (РИНО ПГНИУ)

Цифровая кафедра

Выпускная аттестационная работа

по курсу профессиональной переподготовки «Технологии машинного обучения и нейросети для решения прикладных задач»

**ПРОГНОЗИРОВАНИЕ ИСХОДА МАТЧЕЙ DOTA 2 НА ОСНОВЕ ДАННЫХ О КОМАНДАХ И ИГРОКАХ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ**

Разработчики проекта:

Сафина Полина Наилевна

Останин Егор Федорович

Пермь, 2024

# ПАСПОРТ ПРОЕКТА

**Название проекта:** Прогнозирование исхода матчей Dota 2 на основе данных о командах и игроках с использованием методов машинного обучения.

**Сведения об авторах:** Сафина Полина Наилевна, Останин Егор Федорович.

**Цель:** проведение анализа данных о матчах Dota 2 и построение модели прогнозирования исхода матчей для обеспечения точности предсказаний результатов на основе информации о ко.

**Задачи:**

1. Осуществить загрузку данных и подготовку их к анализу.
2. Провести анализ данных.
3. Выполнить предварительную обработку данных для обучения модели.
4. Обучить модель прогнозирования исходов матчей с использованием методов машинного обучения.
5. Провести оценку качества обученной модели.

**Краткое описание проекта:** в рамках проекта проводится исследование, направленное на разработку модели бинарной классификации. Основной целью данного исследования является выявление факторов, определяющих результаты матчей Dota 2, и создание инструмента, который может быть использован для прогнозирования исходов матчей.

**Конкретные ожидаемые результаты:** модель логистической регрессии, способная эффективно прогнозировать исход матчей.

# СОДЕРЖАНИЕ ПРОЕКТА

## Анализ проблемы исследования

Dota 2 — это популярная многопользовательская онлайн-игра, разработанная компанией Valve, которая является одной из ведущих дисциплин в киберспорте. В каждом матче участвуют две команды — Dire и Radiant, каждая из которых состоит из пяти игроков. Каждый игрок управляет уникальным героем, обладающим особыми способностями, и может носить до шести предметов, влияющих на его эффективность. Исход матчей определяется множеством факторов, включая выбор героев, стратегию команды, действия игроков, взаимодействие героев и использование предметов.

Прогнозирование результатов матчей Dota 2 является сложной задачей из-за большого количества переменных и их взаимодействия. Надежные методы анализа данных становятся всё более востребованными среди игроков, тренеров и аналитиков в условиях растущей популярности киберспорта.

Традиционные подходы к анализу не всегда обеспечивают достаточную точность и требуют значительных временных затрат. Использование методов машинного обучения позволяет выявить скрытые закономерности, которые сложно определить вручную, и создать модель, способную эффективно прогнозировать исходы матчей.

Проект включает обработку данных о матчах, обучение моделеи машинного обучения, анализ значимости факторов и оценку качества модели. Разработка такой системы может значительно повысить точность прогнозирования, улучшить стратегическое планирование команд и предложить новый инструмент для профессионального анализа в сфере киберспорта.

## Исходные данные

В рамках работы использовался датасет "Dota 2 Matches", доступный на платформе Kaggle (<https://www.kaggle.com/datasets/devinanzelmo/dota-2-matches>). Данный датасет включает множество таблиц, содержащих подробную информацию о матчах Dota 2, игроках, героях, предметах и других аспектах игры.

Для анализа и обучения модели были выбраны только следующие таблицы из датасета:

1. match.csv — основная информация о каждом матче, включая идентификаторы и результаты. Всего в таблице содержится 500000 записей.
2. players.csv — данные о каждом игроке в матче, такие как идентификаторы игроков, их героев и используемые предметы.
3. heroes.csv — таблица с информацией о героях.
4. items.csv — таблица с информацией о предметах.

Из указанных таблиц для построения модели машинного обучения были использованы следующие столбцы:

* match\_id — идентификатор матча.
* player\_id — идентификатор игрока.
* hero\_id — идентификатор героя, выбранного игроком.
* item\_0, item\_1, ..., item\_5 — идентификаторы предметов, приобретённых игроком.
* radiant\_win — бинарный показатель результата матча (1 — победа Radiant, 0 — поражение Radiant).

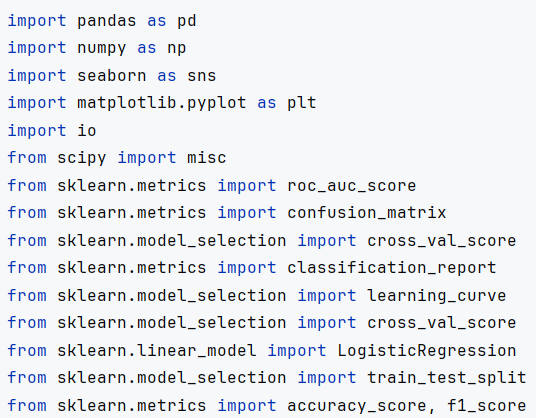
Эти данные позволяют выявить зависимости между действиями игроков, их выбором героев и предметов и исходом матча, что является ключевым для решения задачи прогнозирования.

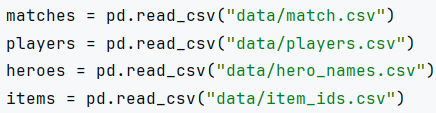
Выдвинем гипотезу: результат матча зависит от состава команд (выбранных героев) и действий игроков (выбранных предметов), и эти данные можно использовать для построения модели машинного обучения, предсказывающей исход матчей.

## Реализация проекта

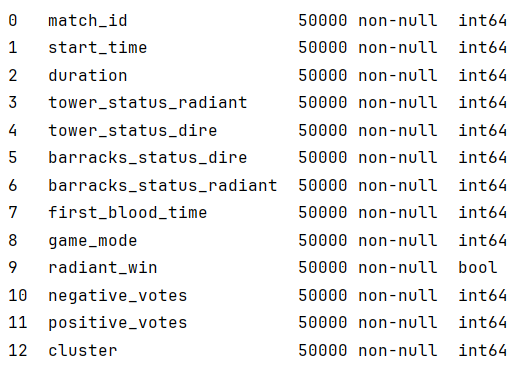
### Этап 1. Подготовка данных к анализу

Загружаем данные и подключаем необходимые библиотеки:



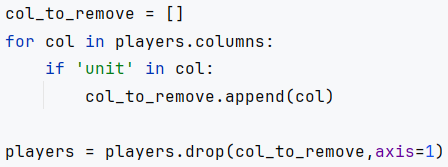


Проверяем типы и нулевые значения в таблице матчей, исключаем избыточные данные:

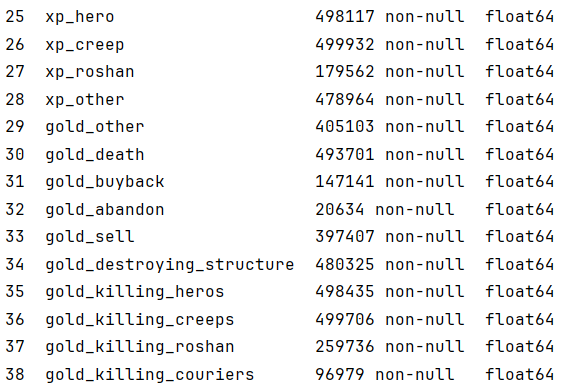




Удаляем столбцы в таблице игроков, начинающиеся на «unit», хранящие избыточную информацию:

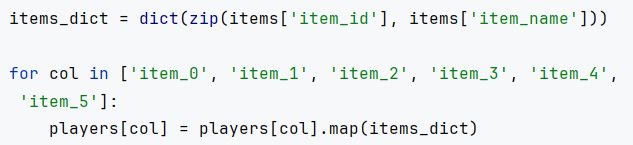


Проверяем наличие нулевых значений в таблице игроков, в данном контексте их можно заменить нулями:

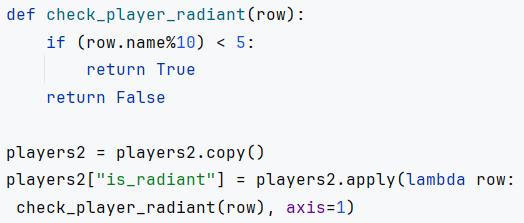




Заменяем id предметов их названиями:



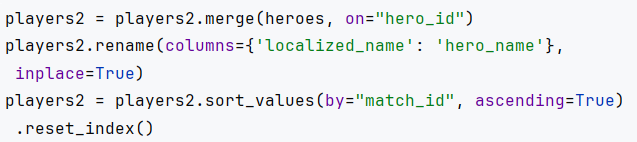
Добавляем столбец, описывающий принадлежит ли игрок команде radiant. Данные в датасете отсортированы по матчам, и первые пять строк матча принадлежат команде radiant, а остальные 5 команде dire:



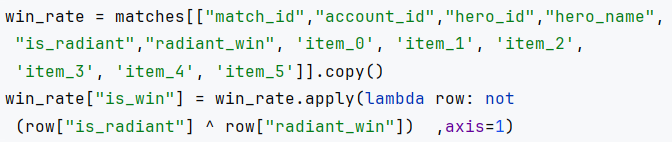
Замечаем, что в таблице с именами героев пропущен id = 24. В таблице игроков также он пропущен, но значения идут с 0, а не с 1 как в таблице с героями. Исправляем эту ошибку увеличив hero\_id в таблице игроков на 1:



Заменяем id героя на его имя:



Выбираем колонки для дальнейшего анализа и создаем новый столбик, описывающий победу или поражение. Для создания этого столбца используем исключающее ИЛИ, которое возвращает True если оба значения столбцов одинаковы:

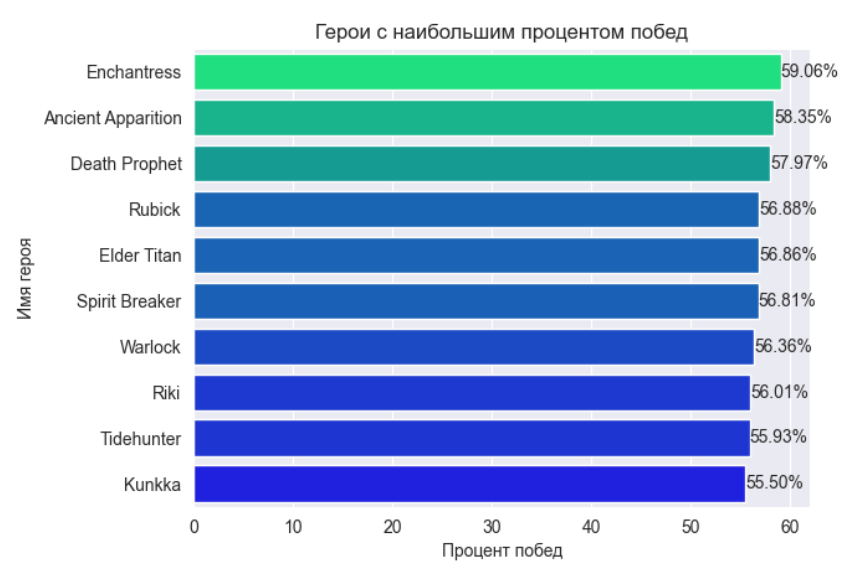


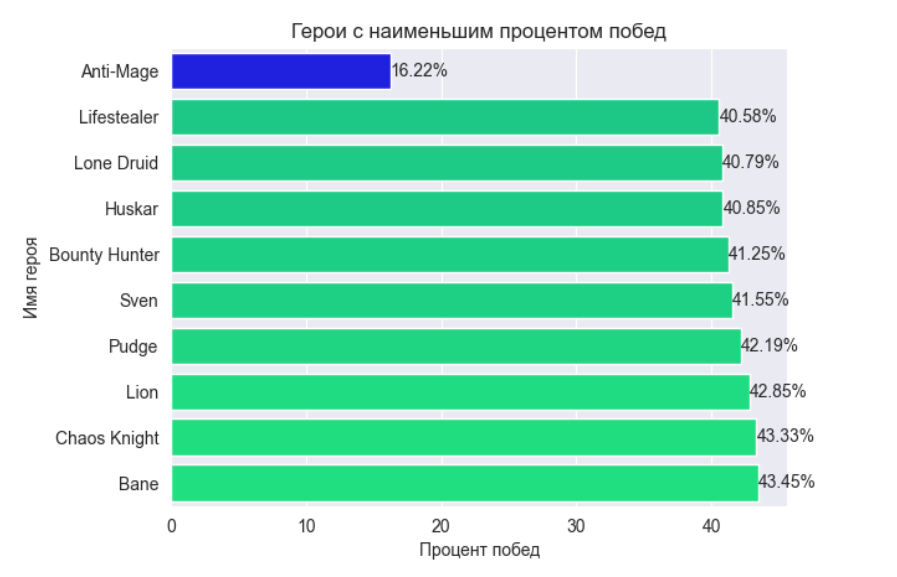
Считаем долю побед для персонажей и предметов:



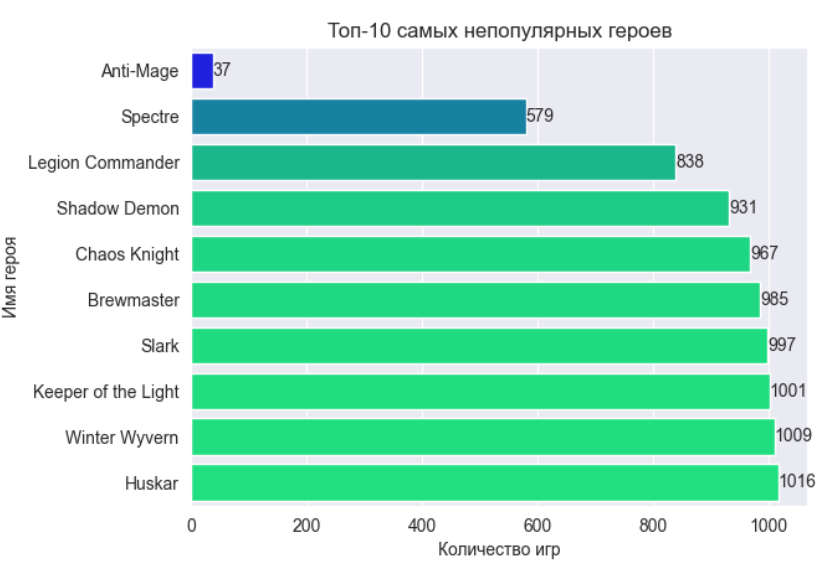
### Этап 2. Анализ данных

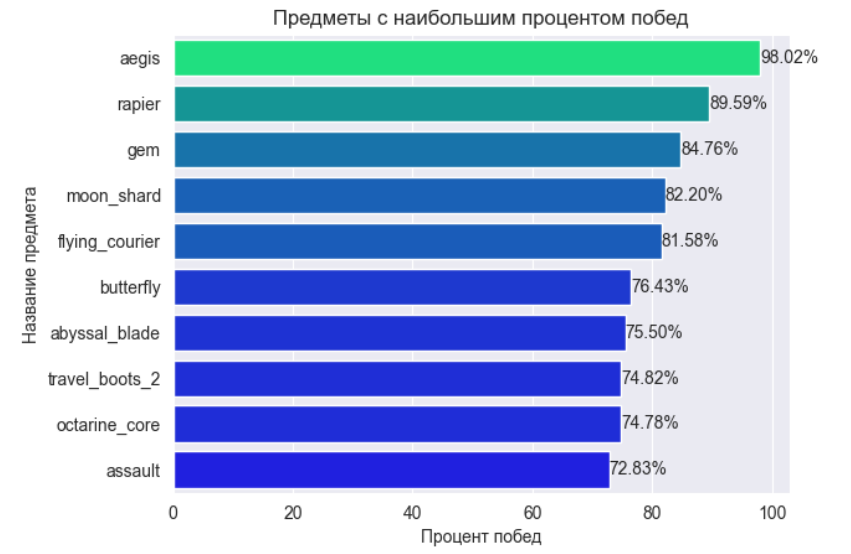
В анализе данных для визуализации используются графики, отображающие героев и предметы с наибольшим и наименьшим процентом побед, а также их популярность. Это позволяет выявить ключевые элементы, оказывающие влияние на исход матчей.

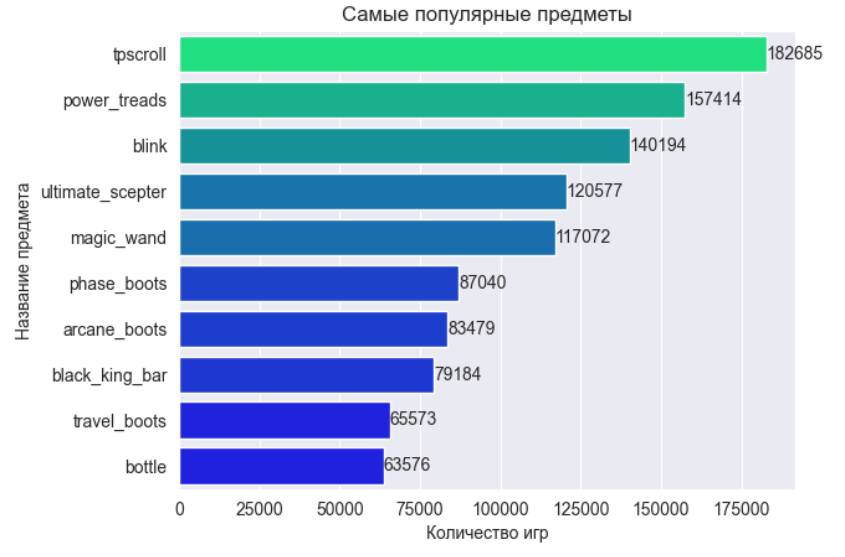








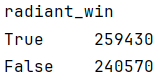
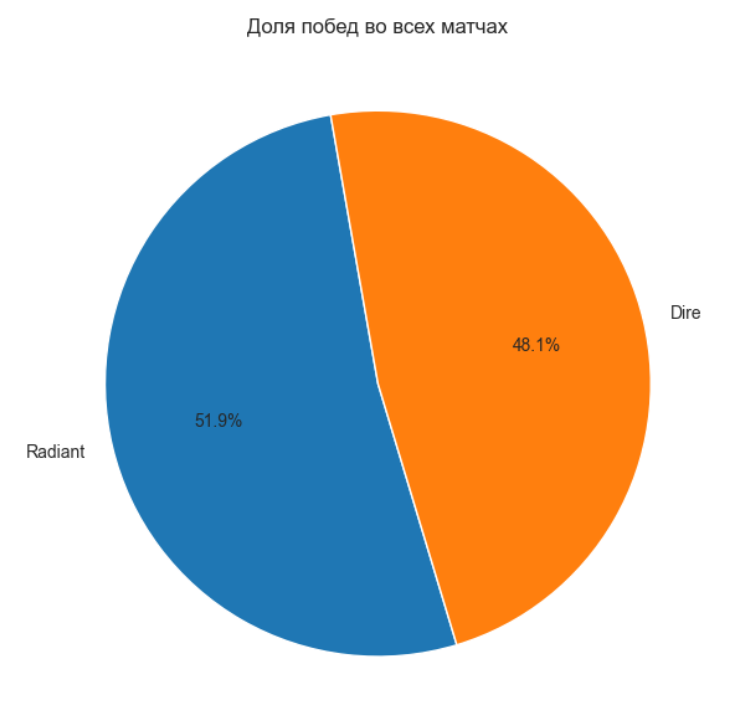




Были выявлены следующие ключевые герои и предметы:

* Герои Enchantress, Ancient Apparition и Death Prophet продемонстрировали наивысший процент побед, свидетельствующий об их высокой эффективности в матчах.
* Среди героев, Zeus оказался самым популярным, в то время как Anti-Mage оказался наименее популярным и с самым низким процентом побед.
* Среди предметов, наибольший процент побед наблюдается у Aegis, Divine Rapier и Gem of True Sight

Посмотрим распределение целевой переменной:



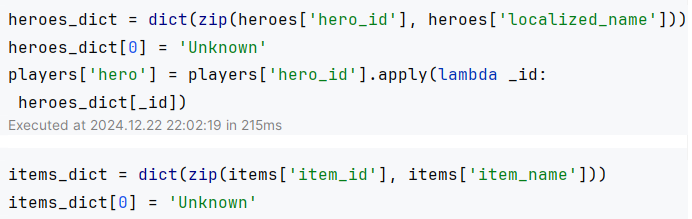
Датасет практически сбалансирован.

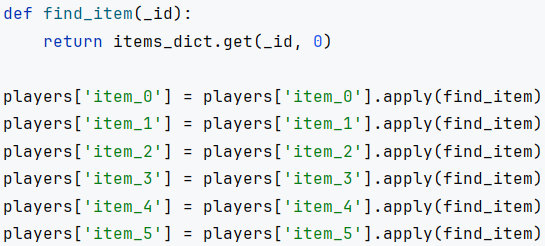
### Этап 3. Подготовка данных к анализу

Выделяем целевую переменную: radiant\_win - победа команды radiant (1 - True, 0 - False):

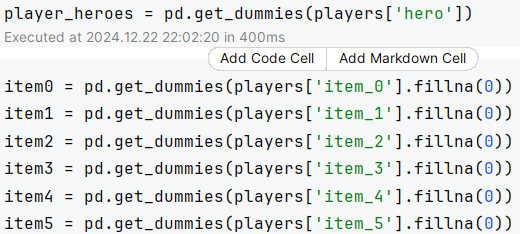


Подготовим признаки для обучения модели. Создаем словари имен героев и названий предметов, подставляем их в таблицу игроков:

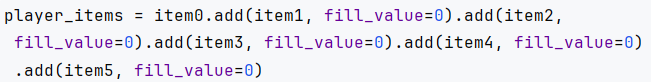




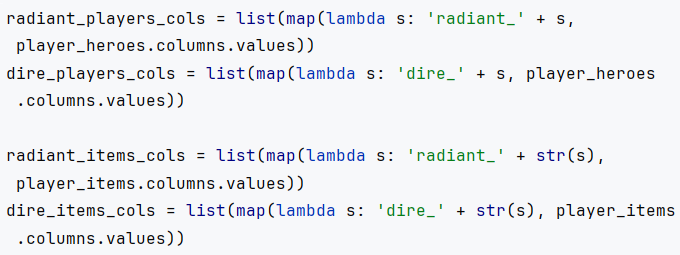
Создаёт новую таблицу, в которой для каждого героя создаётся отдельный столбик с бинарным значением, обозначающий присутствие данного героя. То же самое делаем для предметов:



Так как порядок предметов в данном контексте не важен, объединяем их в одну таблицу, складывая значения:



Подготавливаем колонки для игроков и предметов команды radiant и dire



Группируем данные по матчам и создаем признаки для команд Radiant и Dire, складывая первые 5 строк для Radiant, а следующие 5 для Dire:



Формируем объединенный набор признаков для каждого матча, описывающий количество предметов и героев для каждой из команд:



### Этап 4. Моделирование

Перемешиваем данные для изменения порядка:



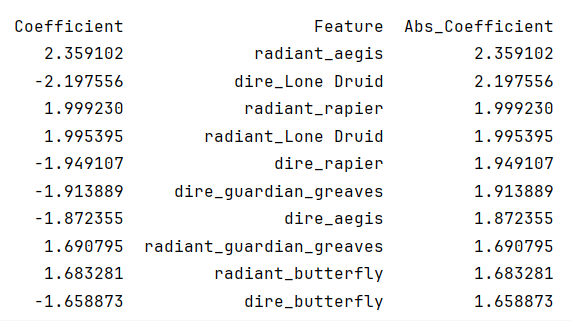
Разделяем X и y на обучающий и тестовый наборы:



Обучаем модель логистической регрессии:

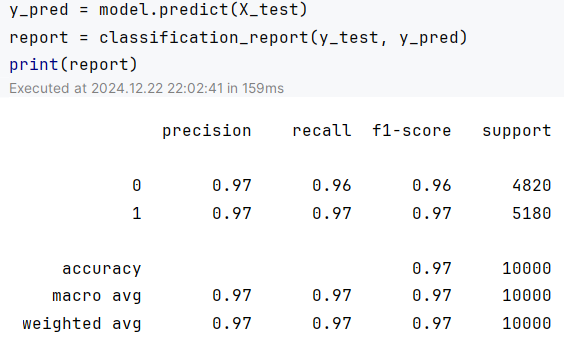


Выведем 10 признаков с наибольшими коэффициентами логистической регрессии, которые отражают влияние признаков на вероятность положительного класса:



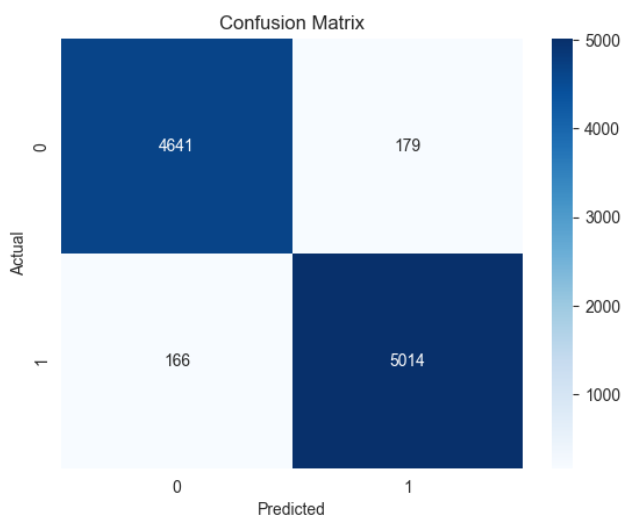
В данном списке присутствует только один герой – Lone Druid. Хотя он не попал в топ героев с наибольшим процентом побед, его влияние на результат матча может быть неочевидным, но значительным.

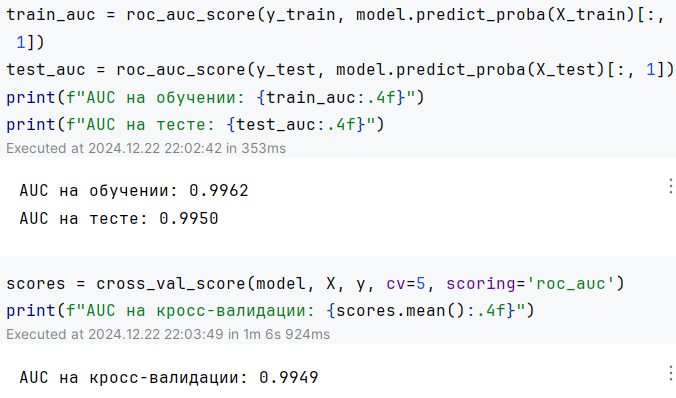
Производим предсказание на тестовом наборе и выводим метрики:



Высокие значения точности и F1-score говорят о том, что модель показывает высокую эффективность в прогнозировании исходов матчей.

Смотрим матрицу ошибок:



  
AUC на обучении, тесте и кросс-валидации имеют высокие, близкие друг к другу, значения, что говорит о хорошей обобщающей способности модели и минимальном переобучении.

# Заключение

В ходе реализации проекта по прогнозированию исходов матчей Dota 2 с использованием методов машинного обучения были выполнены несколько ключевых этапов.

На первом этапе была проведена подготовка данных для анализа. Сначала данные были загружены, после чего проведена обработка пропущенных значений, исключение избыточной информации и объединение данных из разных таблиц по уникальным идентификаторам. Также были созданы новые столбцы, описывающих принадлежность игрока к определённой команде, а также победу игрока в матче. После этого, были вычислены доли побед героев и предметов для последующего анализа.

На втором этапе был проведён анализ данных. С использованием визуализаций были выявлены наиболее значимые герои и предметы, оказывающие влияние на исход матчей. Распределение целевой переменной (победа команды в матче) продемонстрировала сбалансированность датасета.

Третий этап включал подготовку данных для анализа. В первую очередь, была выделена целевая переменная — исход матча (победа Radiant или Dire). После этого, для каждого игрока были созданы бинарные таблицы героев и предметов с использованием метода get\_dummies из библиотеки Pandas. Эти таблицы объединялись по игрокам, затем по командам, а затем — в итоговую таблицу по матчам.

На этапе моделирования была обучена модель машинного обучения — логистическая регрессия. Оценка модели показала высокую точность и стабильность. Логистическая регрессия продемонстрировала высокие значения метрик accuracy — 97% и F1-score — 97%. Значение коэффициента AUC достигло 0.996 на обучении, 0.995 на тесте и 0.994 при кросс-валидации.

Таким образом, использование методов машинного обучения позволяет эффективно предсказывать исходы матчей Dota 2, существенно улучшая точность анализа.

# Приложение

Код программы: https://github.com/elspino/MachineLearningFinal/