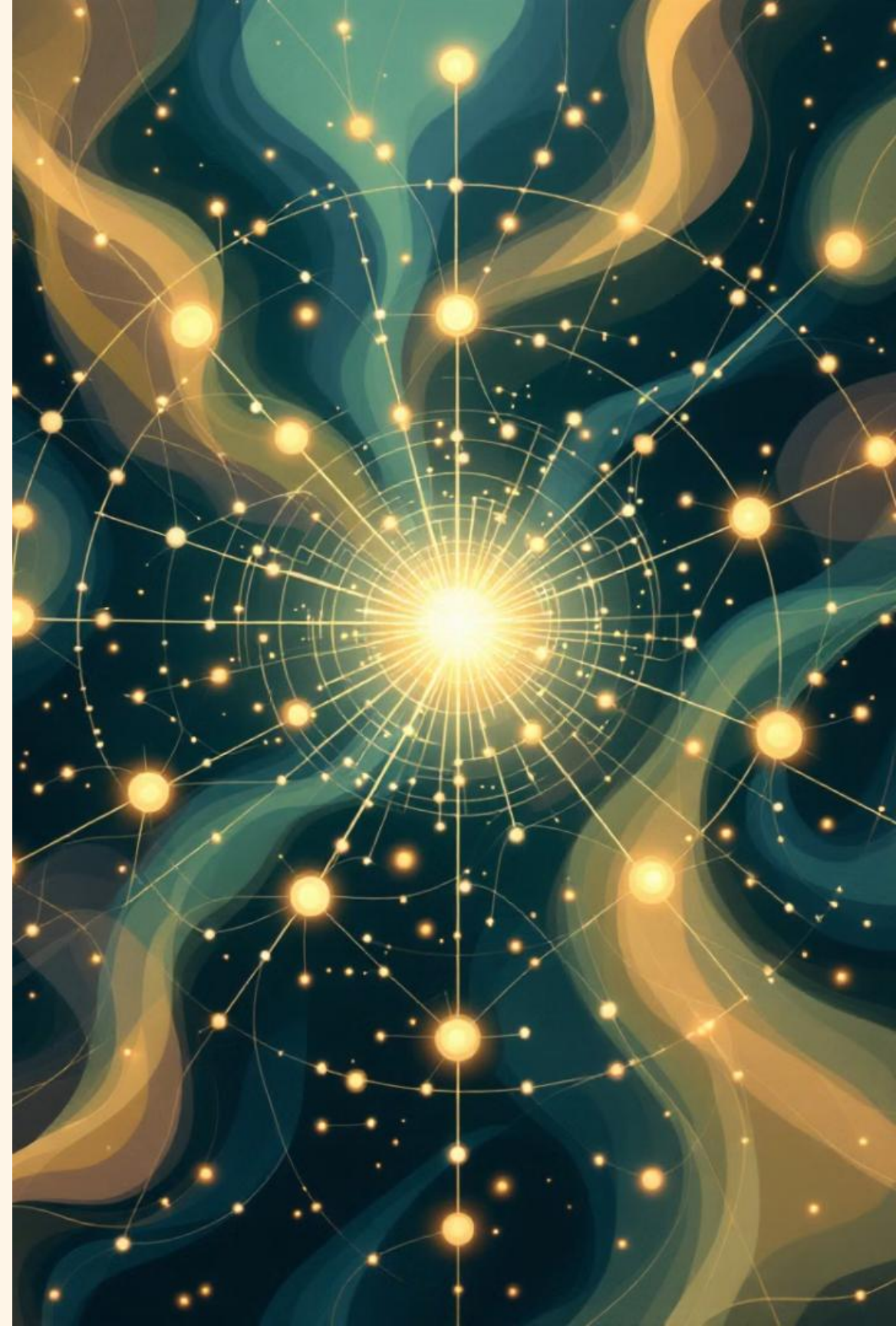


Предсказание Идеального Тарифа для Клиента: Введение

Разработка модели машинного обучения для предсказания оптимального тарифного плана для клиентов телекоммуникационной компании. Цель: Улучшение качества обслуживания клиентов и увеличение их лояльности путем предоставления наиболее подходящих тарифных планов.

 by **Muhammad Olimov**



Кто Я: Мой Путь в Data Science

Имя и фамилия

Олимов Мухаммад

Образование

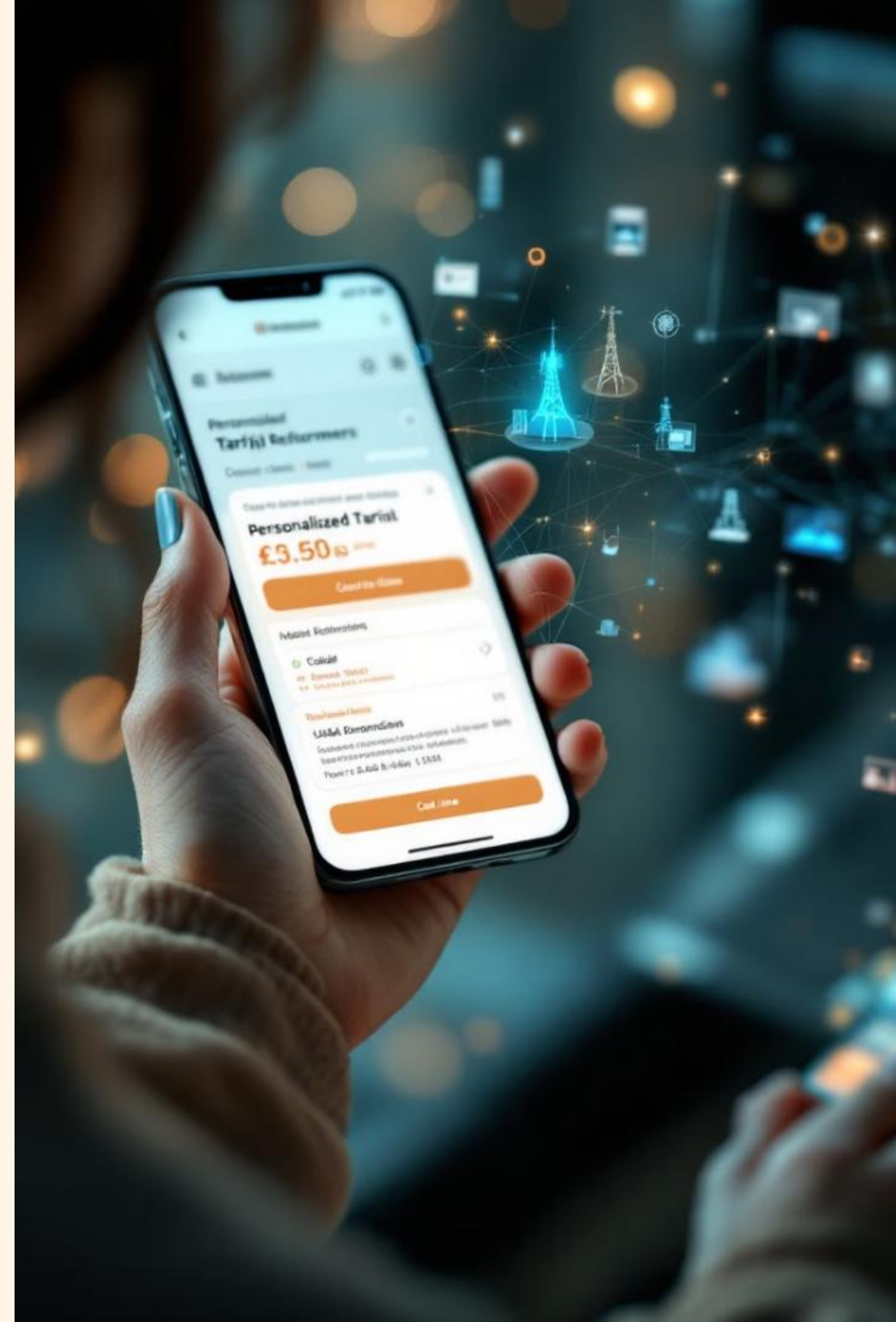
Таджикский технический Университет
имени М.С.Осими,
Автоматизация технических процессов
и производств, 2 курс.

Интересы

Машинное обучение,
анализ данных,
разработка
алгоритмов.

Тема Проекта: Почему Предсказание Тарифов Важно?

- 1 Личная мотивация: Интерес к применению машинного обучения для решения реальных бизнес-задач.
- 2 Польза для меня: Практический опыт работы с данными, улучшение навыков машинного обучения, углубление знаний в области телекоммуникаций.
- 3 Польза для друзей: Возможность рекомендовать оптимальные тарифные планы, экономя их деньги и время.



Бейзлайн: С чего Мы Начинаем?

Бейзлайн модель

Простейшая модель,
KNeighborsClassifier

Точность бейзлайна

Точность для обучающихся
данных: 0.95%.

Точность для тестовых
данных: 0.89%

Недостатки бейзлайна

Не подогнан под
гиперпараметрами.

Не использован более
сложные модели

Метрики Успеха: Оцениваем Качество Модели

Accuracy (Точность)

Доля правильных ответов модели.

Формула: $(TP + TN) / (TP + TN + FP + FN)$, Точность моделья – 80%

Precision (Точность)

Доля правильно предсказанных положительных результатов среди всех предсказанных положительных результатов. Формула: $TP / (TP + FP)$.
Важна, чтобы не предлагать тариф тем, кому он не нужен.

Precision – 83%

Recall (Полнота)

Доля правильно предсказанных положительных результатов среди всех фактических положительных результатов. Формула: $TP / (TP + FN)$.
Важна, чтобы охватить всех, кому тариф может быть полезен.

Recall – 91%

F1-score

Среднее гармоническое между Precision и Recall. Формула:
 $2 * (Precision * Recall) / (Precision + Recall)$.

F1_score – 86%

ROC AUC

Площадь под кривой ошибок. Показывает способность модели различать классы. Значение 0.5 - случайное угадывание, 1 - идеальная модель. ROC-AUC – 95%

Этапы Проекта: Подготовка Данных



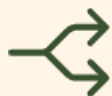
Источник данных:
telco customer churn dataset
из Kaggle



Обработка данных: Я адаптировал целевую переменную на основе нескольких признаков данных, так как не были данные на моем теме



Преобразование данных:
Кодирование категориальных признаков (Target Mean Encoding), масштабирование числовых признаков (StandardScaler).



Разделение данных:
Разделение на обучающую
(80%) и тестовую (20%)
выборки.



Этапы Проекта: Балансировка Данных и Обучение Модели

- 1 — Проблема несбалансированности: Разное количество клиентов, использующих разные тарифные планы. Тариф "Medium" - 4868 клиентов, тариф «Low» - 1526 клиентов и тариф "High" - 649.
- 2 — Методы балансировки: Oversampling (увеличение числа клиентов с менее популярными тарифами)
- 3 — Выбор модели: [RandomForestClassifier], обоснование выбора: (скорость обучения и результат).
- 4 — Обучение модели: Настройка гиперпараметров (GridSearchCV).





Результаты и Выводы

1

Достигнутые метрики: Accuracy = 80%, Precision = 76%, Recall = 85%, F1-score = 76%, ROC AUC = 95%.

2

Выводы: Модель позволяет значительно улучшить предсказание оптимального тарифа по сравнению с простым бейзлайном.

3

Дальнейшие шаги: Улучшение качества данных, использование более сложных моделей, учет дополнительных факторов