Цели и задачи.

Целью задания чемпионата являлась разработка программного обеспечения для предсказания длительности поездки. Для этого мне было предложено провести исследование данных, подаваемых на вход моего решения, после чего дополнить эти данные новыми признаками из дополнительного набора и сконструировав их на основе существующих. На разработанных мною данных мне было необходимо создать модель для предсказания длительности поездки и сконструировать приложение, которое позволит пользователю с удобством взаимодействовать с моделью.

Сессия 1.

На данном этапе, моей задачей был первичный анализ данных и дополнение данных с использованием дополнительного датасета.

В самом начале мне было необходимо объединить 2 части обучающей выборки, после чего удалить дубликаты строк, обработать пропущенные значения и привести данные в единый формат.

После этого , с использование дополнительного набора, я добавил новые признаки, а именно: **указать какие.** Итоговый датасет после данных преобразования выглядит следующим образом (**рисунок 1).**

Сессия 2.

После первичного анализа данных, мне было предложено разработать дополнительные признаки, которые могли бы улучшить качество модели в будущем.

В качестве дополнительных признаков мною были выбраны следующие временные признаки: сезон поездки, день недели и время дня или ночи. Данные признаки могут повлиять на модель, т.к. Они связаны с длительностью поездки (например в ночное время низкая загруженность дорог и длительность в среднем должна быть меньше чем в дневное время) **графики зависимости целевой переменной и данных признаков приведены на рисунке 2**.

В качестве гео-признака, я добавил расстояние поездки. Выполнил этот признак я, использовав координаты начала и конца поездки. При расчете расстояния я использовал формулу на **рисунке 3,** которая учитывает кривизну земного шара.

Помимо этого, с использованием координат начала и конца поездки, мною был разработан новый признак с использованием кластеризации зоны поездки - Район поездки.

После добавления признаков я провел анализ важности признаков, используя метод основных компонентов и график важности признаков я выбрал признаки необходимые мне для модели: **НАПИСАТЬ КАКИЕ**

Сессия 3.

В результате подготовки данных, я получил очищенный дополненный набор данных, который будет подаваться на вход модели. На данном этапе моей задачи было исследовать и найти оптимальный алгоритм машинного обучения, оптимизировать его и обучить модель. Основываясь на линейную зависимость между переменными, в качестве базового алгоритма я выбрал линейную регрессию. Данный алгоритм показал неплохой результат на основном и дополненном новыми данными датасете.

После чего, мною были исследованы следующие алгоритмы: KNeighboursRegressor, DecisionTreeRegressor, RandomForestRegressor. Результат сравнения с использованием кросс-валидации приведен на графике на **рисунке 4.** Как видно из графика, линейная регрессия показала наилучший результат. Помимо качества модели данная модель быстро обучается и может улучшаться использованием дополнительных алгоритмов, таких как Lasso и Ridge.

Я также выполнил оптимизацию модели с использованием оптимизации гипер-параметров и масштабированием данных.

Сессия 4.

Последним шагом моей работы была разработка прикладного программного обеспечения для пользователя, который будет взаимодействовать с моделью.

На данном этапе мною был разработан конвейер трансформации, который использовался разработанным ПО для преобразования подаваемых пользователем данных. Таким образом, логика приложения не будет пересекаться с логикой обработки данных.

Для разработки графического интерфейса я использовал графический модуль python - TKinter. Данный модуль прост в использовании и позволяет выполнять в приложении необходимый набор операций. Итоговый результат интерфейса приведен на **рисунке 5.**

**Скопировать описание интерфейса из руководства**