Порядок решения классических ML задач (примерный):

1. Обработка данных:
   1. Открыть датасет;
   2. Смотрим на его характеристики (какие есть фичи (обычно есть описание), какие из них категориальные, есть ли пропуски и сколько, есть ли разность в масштабах);
   3. Удаляем / заполняем пропуски;
   4. Масштабируем вещественные признаки. Иногда нужно масштабировать и отклик.
   5. Обрабатываем категориальные признаки;
2. Feature Engineering:
   1. Определить ненужные признаки и удалить их (корреляционный анализ, диаграммы рассеяния, fstr моделей).
   2. Выделить новые признаки (возможно, из категориальных необработанных на прошлом этапе);
3. Обучение модели:
   1. Разделяется тест (просто некоторая отложенная выборка) и трэйн. Есть еще реальный тест задачи, для которого ничего не известно. Не забываем про стратификацию.
   2. Выбираем модель (Регресси: LinearRegression, KNN, DecisionTree, Random Forest, catboost, xgboost, Ridge, Lasso, MLP; Класификация: все то же самое, но LinearRegression -> LogisticRegression + SVM).
   3. Выбираем метрику и подбираем гиперпараметры на кросс-валидация на трейне. Находим модель с наилучшими гиперпараметрами. ЕСЛИ ВРЕМЕНИ МАЛО, то трейн можно разбить на трейн и валидацию один раз и проверять так каждую модель (без кросс-валидации).
   4. Проверяем теперь лучшую модель на тесте через fit, predict. Если устраивает (качество не сильно упало), то считаем, что все готово. Иначе учимся заново.
4. Отправка резултата:
   1. МОЖНО, но не обязательно обучиться на ВСЕМ наборе предоставленных данных.
   2. Берем тест соревнования и проделываем с ним то же самое, что и с трейном, делаем предсказания, формируем файл и отправляем в систему.