Playbook Data Hack

Vorbemerkung

Ist das "Template" für den Data Hack (wird eine Kaggle Competition sein und am Abend gibt es eine submission). Hier hangeln wir uns dann am Datahack entlang und können es brauchen, falls wir nicht mehr weiter wissen. Wichtig ist auch zu verstehen, dass die Accuracy nicht das zwingende Hauptziel ist, sondern auch der Erkenntnisgewinn aus dem Prozess. Als Vorbereitung für den Data Hack dann auch schon mal eine kaggle challenge durchspielen (oder zumindest in Teilen).

Vorgehen

1. Dataset der Challenge herunterladen
2. Feature Analysis / Feature Engineering
    - Daten einlesen via Pandas, Dataframe erstellen
        - Visualisieren, darstellen, z.B. Pair plot, Korrelationen, min/max, Wertebereich
    - Datenzustand prüfen
        - Complete/Incomplete, z.B. null values
        - Spaltenverwechslungen, z.B, sanity check, Datentypen (z.B. wenn Spalte numerische sein sollte aber nicht-numerische Werte drin sind, dtype)
        - Outlier prüfen, z.B. DB-Scan, Abweichungen +/- 3 σ
    - Bereinigen der Daten (löschen)
    - Ziel: Sauberes und strukturiertes Dataset
3. Modellphase
    - Splitten Train/Val oder Folding (cross-validation, KFold)
    - Datenmanipulation des Trainingsets
        - Standardisieren
        - Löschen
        - Transformieren
        - Skalieren (min/max, logarithmisch,…)
        - Eventuelle neue Features ausrechnen, feature composition oder decomposition

- Problemstellung
    - Classification (Hund oder Katze) oder Regression (hauspreis)
    - Modellauswahl (Forrest, SVM, xgboost,…) - abhängig von der Art der Problemstellung (mit grosser Wahrscheinlichkeit wird es xgboost sein 😉)
- Test- und Training Loop
    - Trainingscore: Resultate des Outputs prüfen
    - Validation Score: Modelqualität prüfen, Erkennen eventuelles Overfitting/Underfitting -> Unbedingt overfitting vermeiden
    - Für Outliers die Labels anschauen. Wenn also das Modell mit grosser Wahrscheinlichkeit daneben liegt. Und dann schauen, ob die Rohdaten falsch sind/falsch gelabelt sind und dann allenfalls eine Reclassification machen). Hinweis: Für Kaggle nicht so spannend/korrekt, wohl aber für das echte Leben.
    - Features variieren, experimentieren; kann ich Features entfernen und so
    - Hyperparameter tunen - nicht nur brute force, sondern auch mit Überlegung