2024

Problematic Internet Use

INTRO TO DATA SCIENCE PROJECT

AHMED WAEL FOUAD SOLIMAN – A3

Setting Up the Project

https://github.com/ahmedwael19/Prediction-of-Internet-Use-Kaggle-Compeititon.git

Business Understanding

Business Goals Identification

Background

Problematic internet use is an increasingly critical issue among children and adolescents. This overuse has been linked to severe mental health challenges, including depression and anxiety. Despite its popularity, assessing internet usage problems typically require specialized tools and expertise, often inaccessible to many families. Consequently, such assessments are rarely performed, often leaving the problem unchecked.

To address this issue, the **Child Mind Institute**—an independent nonprofit supporting children with mental health and learning disorders—collaborated with **NVIDIA** and **Dell**. Their initiative proposes using easily measurable physical fitness indicators as proxies for detecting problematic internet use. This novel approach can be beneficial in environments where clinical expertise or traditional assessment tools are unavailable.

As part of this collaboration, the **Kaggle Competition** challenges participants to develop a predictive model capable of analyzing children's physical activity and fitness data to identify early signs of problematic internet use. This initiative aims to promote timely interventions and encourage healthier digital habits in children and adolescents.

Business Goals

1. Combine and analyze the datasets offered by the competition and extract actionable insights through data visualization.
2. Develop a Machine Learning (ML) model that can achieve a top 15% ranking position within the Kaggle competition public leaderboard

Business success criteria

1. Extract at least four key insights from the datasets.
2. Achieve a position in the top 15% Kaggle competition public leaderboard.

Situation Assessment

Inventory of resources

Hardware and Infrastructure

- **Laptop**: HP EliteBook 640, 16 GB RAM, Intel Core i5-125U processor.

Data Resources

- Datasets provided by the Mind Child Institute Competition on Kaggle.

Software Tools

- PyCharm Professional – student license-
- Anaconda Distribution
- Google Colab Jupyter notebooks
- GitHub Version control
- Kaggle Kernels

Requirements, Assumptions, and Constraints

Requirements

- Thorough data cleaning, with clear and valid justifications for each preprocessing step.
- Interpretable and explainable feature engineering.
- Actionable insights must be derived from non-trivial data, which cannot be directly inferred from raw data.
- The ML model must outperform a random classifier.

Assumptions

- The public leaderboard reflects the model's success, as the competition concludes after the project deadline.
- The dataset remains accessible on Kaggle throughout the project's duration.

Constraints

- The project must be completed before the competition ends.

Risks and contingencies

Technical Risks

Risk: Out-of-memory errors

Contingency: Use data chunking to fit data into memory.

Risk: Google Colab server becomes unavailable.

Contingency: Maintain offline backups and local versions of the code.

Risk: Colab session times out after 6 hours.

Contingency: Ensure experiments and trials fit within the time limit.

Data Risks

Risk: Datasets become inaccessible.

Contingency: Maintain periodic backups

Implementation Risks

Risk: ML Model underperforms

Contingency: Experiment with diverse ML models (from simple to complex) and apply hyperparameter tuning

Terminology

- **quadratic weighted kappa**: A metric to evaluate agreement between predictions and outcomes. Scores range from 0 (random agreement) to 1 (perfect agreement). Negative scores indicate less agreement than random.
- **SSI (Severity Impairment Index):** The target outcome represents the level of problematic internet use in children.

Costs and benefits

Costs

There are no financial costs as all required resources (hardware, software, and data) are freely available.

Benefits

- Develop an evidence-based understanding of the relationship between physical activity and problematic internet use, providing insights for parents and policymakers.
- Build a robust ML model for detecting early signs of internet addiction, potentially leading to targeted intervention strategies and recovery plans for affected individuals.

Data-Mining Goals

Goals

This project's overarching goal of data mining is to transform raw data into actionable insights and predictive capabilities. Specifically, the objectives include:

1. Data Exploration and Preparation
    - Explore the provided dataset to understand its structure, key features, and distribution.

- Perform thorough data cleaning, including handling missing values, outliers, and irrelevant features.

2. Feature Engineering
    - Create meaningful and interpretable features from the raw dataset that correlate strongly with the target variable (Severity Impairment Index - SSI).
    - Identify physical activity patterns that serve as proxies for problematic internet use.
3. Model Development and Evaluation
    - Develop a machine learning model that accurately predicts SSI levels based on the data.
    - Ensure the model generalizes well using robust evaluation techniques like cross-validation.
4. Visualization and Insight Generation
    - Generate at least four actionable insights from the dataset that help understand the relationship between physical activity and internet addiction.
    - Provide visualizations to illustrate these insights and support decision-making.

Success Criteria

The success of the data-mining process will be evaluated based on the following criteria:

1. Exploratory Data Analysis (EDA)
    - Identify at least four valuable insights that are non-obvious and offer an actionable understanding of problematic internet use.

2. Feature Engineering
    - Successfully create features that are both interpretable and have strong predictive power, ensuring transparency for stakeholders like parents and educators.
3. Model Performance
    - Achieve a Quadratic Weighted Kappa score high enough to place within the top 15% of the Kaggle competition leaderboard.
4. Visualization Quality
    - Develop clear, engaging, and professional visualizations that effectively communicate insights and patterns in the data.

Data understanding

Gathering Data

Data Requirements

To successfully gain actionable insights and predict the Internet Addiction Index (SSI), the following datasets are required:

HBN Instruments

- **Description**: Tabular data containing various psychological, physical, and behavioral measurements from instruments such as demographic surveys, physical measures, internet usage data, and more.
- **Purpose**: Serves as the primary dataset, offering a wide range of variables essential for understanding participant characteristics, behaviors, and their relationship with problematic internet use.

Actigraphy Files

- **Description**: While participating in the study, some participants were given an accelerometer to wear for up to 30 days while at home and going about their daily lives. This data is available in parquet format, partitioned by participant ID.
- **Purpose:** Provides detailed physical activity data, including acceleration and motion metrics, which can help identify patterns correlating with problematic internet use.

Data Availability Verification

HBN Instruments

Fully available in the Kaggle competition for all **3,960**

Actigraphy Files

Available for only **996 participants**.

Selection Criteria Definition

HBN Instruments

- **Files**: *train.csv and test.csv*
- **Relevant Fields**:
  - id: Participant identifier.
  - Demographics: Age, sex.
  - Physical Activity Questionnaire: Vigorous activity indicators.
  - Sleep Disturbance Scale: Sleep disorder categories.
  - Parent-Child Internet Addiction Test (PCIAT): Internet addiction severity.
  - Physical Measures: Height, weight, BMI, blood pressure.
- **Case Range**: Data for all 3,960 participants.

Actigraphy Data

- **Files**: *series_{train|test}.parquet*
- **Relevant Fields**:
  - id: Participant identifier.
  - X, Y, Z: Acceleration in g-force along each axis.
  - enmo: Euclidean Norm Minus One of the acceleration signals.
  - anglez: Arm angle relative to the horizontal plane.
  - time_of_day: Time of each 5-second interval.
  - non-wear_flag: Indicates whether the device was worn.
- **Case Range**: Subset of 996 participants with accelerometer data.

Cross Analysis

Merging Criterion: Combine datasets using the id field to align participant data between train.csv/test.csv and actigraphy data.

Describing Data

HBN Instruments (CSV Files)

- **Source**: train.csv and test.csv
- **Format**: Tabular CSV files.
- **Number of Cases**: Data for 3,960 participants.
- **Fields**: Includes various psychological, physical, and behavioral measures such as demographics, internet use, physical activity, sleep disorders, and body composition. Field descriptions are provided in data_dictionary.csv.

Actigraphy Data (Parquet Files)

- **Source**: series_{train|test}.parquet

- **Format**: Parquet files for continuous time-series accelerometer data.
- **Number of Cases**: Data available for 996 participants.
- **Fields**:
    - id: Unique patient identifier.
    - step: Time step for each observation.
    - X, Y, Z: Acceleration in g-force along three spatial axes.
    - enmo: Euclidean Norm Minus One, representing motion magnitude.
    - anglez: Angle of the arm relative to the horizontal plane.
    - non-wear_flag: Indicates whether the device was being worn (0: worn, 1: not worn).
    - light: Ambient light intensity in lux.
    - battery_voltage: Device battery level in mV.
    - time_of_day: Timestamp for each sampled interval.
    - weekday: Day of the week (1: Monday, 7: Sunday).
    - quarter: Quarter of the year (1 to 4).
    - relative_date_PCIAT: Days relative to the administration of the PCIAT test.

Suitability of the Data

Field Coverage

The datasets include critical fields for predicting the Internet Addiction Index (SSI), such as physical activity measures, demographics, and internet use behavior.

Case Volume

The datasets provide sufficient cases (3,960 for HBN Instruments and 996 for Actigraphy Data) to develop and validate predictive models.

Data Quality Challenges

- **Missing Values**: Present in both datasets; will require imputation strategies.
- **Imbalance**: The actigraphy data covers only a subset of participants, which may affect combined analyses.
- **Too many features**: L1 Regularization, Variance Thresholding, or a similar strategy will remove features with very low variance, as they are unlikely to contribute meaningfully to the model.
- **Variability**: Actigraphy data is a time series requiring consistent resampling. Some rows might be resampled or aggregated since we have measurements every 5 seconds for each patient in different time windows (seconds -> minutes)
- **Missing Predictions**: Predicting values are missing in some rows, so we will impute them based on KNN, clustering, or simple imputation using the mean or median.

Exploring Data

HBN Instruments

General Observations

- Contains 82 columns with diverse metrics ranging from demographics to fitness and psychological assessments.
- Data Types
  - 68 columns are numeric (float64 or int64).
  - 12 columns are categorical (object).
- **Target Variable (sii)**: Severity Impairment Index has 1,224 missing entries (~31% missing), which may affect the modeling process

Key Insights

- **Demographics**:

  - Basic_Demos-Age: Mean age is ~10.4 years, ranging from 5 to 22 years.
  - Basic_Demos-Sex: Binary representation (0: Male, 1: Female) with a fairly balanced distribution.
- **Physical Measures**:
  - Physical-BMI: Mean BMI is ~19.3, but there are outliers (values up to 59.13).
  - Physical-Height and Physical-Weight: Outliers in both height and weight, with max weight reaching 315 kg.
- **PCIAT (Internet Addiction Test)**:
  - Most individual PCIAT scores (PCIAT-01 to PCIAT-20) are well-covered (2733–2736 non-null entries).
  - PCIAT_Total: Total score has a mean of ~27.89, with a wide range (17 to 93).
- **Missing Data**:
  - **Significant missing values** in columns like Physical-Waist_Circumference (77% missing) and PAQ_A-PAQ_A_Total (88% missing).
  - Psychological scales like CGAS-CGAS_Score and SDS-SDS_Total_Raw have ~40% missing values

Actigraphy Data

General Observations

- The dataset comprises time-series data for 996 unique IDs, each with 13 columns (X, Y, Z, enmo, anglez, non-wear_flag, light, battery_voltage, time_of_day, weekday, quarter, relative_date_PCIAT, and step).

Preprocessed Features

(Stat_* features) are derived from the original columns -other than step- for each ID, capturing aggregate descriptors:

- Count
- Mean
- Standard Deviation (std)
- Minimum (min)
- 25th Percentile (25%)
- Median (50%)
- 75th Percentile (75%)
- Maximum (max)

So, in total, we have 96 features, which are all numeric and do not have missing values.

Key Insights

- **Statistical Measures:**
    - Features like *Stat_76, Stat_47*, and *Stat_9* exhibit wide ranges, indicating participant variability.
    - Extreme values are observed in metrics such as:
        - *Stat_40*: Range from -90 to -79.8
        - Stat_80: Extremely high values, indicating potential outliers.
    - Features like *Stat_88* and *Stat_85* have distributions centered near the mean, with minimal skewness.
- **Distributions:**
    - Most features follow normal distributions or are moderately skewed.
    - Features like *Stat_0* and *Stat_65* show heavy skewness.
- **Outliers:**
    - Extreme values observed in *Stat_20, Stat_92,* and *Stat_36* may represent rare participant behavior or anomalies.
- **Raw Data:**
    - This dataset summarizes the underlying time-series data, which contains additional granularity and may reveal temporal relationships not captured here.

Verifying Data Quality

HBN Instruments

Strengths

- Comprehensive coverage of demographic and psychological metrics.
- Includes the target variable (sii) and key predictors like PCIAT scores.

Issues

1. **High Missingness**:
   - Critical features like CGAS-CGAS_Score (~40%) have substantial missingness.
   - The target feature, sii, has 31% missing values
   - Features with >75% missing values (e.g., PAQ_A-*, Fitness_Endurance-*) are likely unusable.
2. **Outliers**:
   - Extreme values in physical measures may skew models.
3. **Feature Redundancy**:
   - Correlation analysis is needed to eliminate redundant or irrelevant features.

Action Plan

- Impute missing values using median, clustering, or KNN for key predictors.
- Drop features with excessive missingness (>75%) unless critical.
- Apply z-score or IQR filtering to address outliers.

Actigraphy Data

Strengths

- **Preprocessed Aggregation**:

    - Summaries like mean, variance, min, and max for each variable allow for scalable and efficient analysis.
    - No explicit missing values in the aggregated dataset.
- **Wide Feature Set**:
    - 92 features comprehensively represent sensor metrics.

Issues

1. **Loss of Temporal Granularity**:
    - Aggregation removes time-series relationships, which may limit the ability to study behavioral patterns over time.
2. **Outliers**:
    - Features like Stat_80 and Stat_36 exhibit extreme ranges that may skew results.
3. **Feature Redundancy**:
    - Several features likely overlap (e.g., statistical moments of the same variable), requiring correlation analysis and feature selection.

Action Plan

- **For Aggregated Data**:
    - Normalize features with extreme values to reduce skew.
    - Conduct correlation analysis to identify redundant features.
    - Perform clustering to detect participants with unusual behavior.
- **For Raw Data**:
    - Validate aggregated statistics against the raw time-series data.
    - Analyze temporal patterns in the raw data to explore trends and anomalies.

- **Modeling Consideration**:
  - Start with the aggregated dataset for exploratory analysis and initial modeling.
  - If necessary, return to raw data for temporal analysis or alternative feature engineering.

Planning

Tasks

All the tasks are done by Ahmed Wael, as this is a solo project.

1. Business Understanding (4 hours)

- Understand objectives, evaluation metrics, and expectations and read competition details. (2 hours)
- Research actigraphy data and its applications in behavioral health and machine learning. (2 hours)

2. Data Collection and Acquisition (25 hours)

HBN Instruments Dataset (9 hours)

- Initial data collection and cleaning (3 hours).
- Identify and impute missing values using median and KNN-based imputation for categorical and numerical features (3 hours).
- Normalization and label encoding for categorical columns (3 hours).

Actigraphy Dataset (16 hours)

- **Aggregated Data (10 hours)**
  - Initial data loading and aggregation into statistical summaries (mean, variance, min, max) for all numeric columns using PySpark (3 hours).

    - Checking aggregated statistics against raw data for consistency (4 hours).
    - Handling potential outliers in the aggregated statistics (e.g., extreme values in Stat_80 and Stat_36) by normalization and scaling (3 hours).
- **Raw Data (6 hours)**
    - Temporal pattern exploration to identify trends, anomalies, and seasonal components using Prophet (3 hours).
    - Analyze raw time-series data for potential alternative feature extraction (3 hours).

3. Exploratory Data Analysis (EDA) and Feature Engineering (18 hours)

Actigraphy Aggregated Dataset (6 hours)

- Statistical exploration of aggregated features (e.g., distributions, outliers, correlations) (3 hours).
- Correlation analysis to identify and reduce redundant features (e.g., overlapping statistical moments) (3 hours).

HBN Dataset (6 hours)

- Analyze distributions of variables like demographic metrics and PCIAT scores. (3 hours)
- Investigate missing value patterns to ensure data quality (3 hours)

Feature Engineering (6 hours)

- **Actigraphy Aggregated Dataset:**
    - Clustering participants to detect unusual behaviors (3 hours).
- **HBN Dataset:**
    - Use t-SNE for dimensionality reduction to identify significant features (3 hours).

4. Dataset Integration and Cross-Analysis (5 hours)

- Merge datasets on ID using PySpark to handle memory constraints (2 hours).
- Analyze correlations and interactions between features from both datasets (3 hours).

5. Modeling and Evaluation (30 hours)

Model Training (20 hours)

- Train baseline models (XGBoost, Random Forest, Logistic Regression) using aggregated actigraphy data and the HBN dataset. (10 hours)
- Analyze the importance of features in refining predictors (5 hours).
- Experiment with advanced preprocessing and feature selection (5 hours).

Hyperparameter Tuning and Validation (7 hours)

- Automate tuning tools like Optuna or Hyperopt for hyperparameter optimization.
- Conduct cross-validation to ensure model robustness.

Model Evaluation (3 hours)

- Evaluate models using accuracy, F1-score, ROC-AUC, and other classification metrics.
- Benchmark results against project objectives.

6. Documentation and Reporting (12 hours)

- Compile a comprehensive report with all findings, methods, and results (4 hours).
- Prepare a poster to present the key findings (8 hours).

Total Hours

84 hours

Tools and Software

- PySpark
- Pandas
- NumPy
- Prophet
- XGBoost
- Sklearn
- Seaborn
- Statsmodel
- Optuna
- Hyperopt
- Auto-sklearn