Project E9: KAGGLE - STUDENTS' PERFORMANCE ANALYSIS

Team members : Annabel Aleksius, Mattias Minejev, Tõnis Tõnissoo

Business Understanding Report

Identifying Your Business Goals

Background

High school academic performance forms the basis of opportunities that students will have in higher education and beyond. However, a student's grades are heavily influenced by several factors, including study habits, parental involvement, and extracurricular activities. In analyzing this dataset, we seek to gain insight into these factors and further develop a predictive model to identify students who might be at risk of poor performance, enabling timely intervention.

Business Goals

- Identify key factors that influence students' academic performance.
- Develop a predictive model that categorizes the students' grades into the following categories: 'A', 'B', 'C', 'D', 'F'.
- Provide actionable insights for educators and parents for targeted interventions.

Business Success Criteria

- The model must achieve an accuracy of at least 85% to predict the GradeClass variable.
- Provide interpretable results highlighting the top factors affecting performance.
- Based on the models developed, create a set of recommendations that can improve the academic outcomes of the students falling in grades 'C', 'D', and 'F' by at least 10%.

Assessing Your Situation

Resources Inventory

- Dataset: A synthetic dataset of 2,392 high school students with attributes such as demographics, study habits, parental involvement, extracurricular activities, and GPA.
- Tools and Technology: Python (for data preprocessing and modeling), machine learning libraries (e.g., Scikit-learn, TensorFlow), and visualization platforms (e.g., Tableau or Matplotlib).
- Human Resources: Data scientists, educators, and educational policy advisors.
- Infrastructure: Cloud-based resources for data storage and model training.

Requirements, Assumptions, and Constraints

- Requirements:
  - Access to all features in the dataset, particularly GPA, GradeClass, and other influencing factors like StudyTimeWeekly and ParentalSupport.
  - Collaboration with educational experts to validate insights.
- Assumptions:
  - The dataset accurately represents typical high school students' demographics and behaviors.
  - External factors like curriculum changes or policy variations do not significantly alter outcomes during the study.
- Constraints:
  - Limited time (three months) to complete analysis and modeling.
  - The synthetic nature of the dataset may limit generalizability to real-world applications.

Risks and Contingencies

- Risk: Data may be missing or biased, for instance, not having enough data for one ethnicity.
- Mitigation: Impute missing data and perform checks for bias in EDA.
- Risk: Model overfitting with a small/synthetic dataset.
- Mitigation: Perform robust cross-validation techniques, emphasize generalizability of the model.
- Risk: Educators might resist the recommendations. Mitigation: Communicate results in a digestible format, involve stakeholders early on.

Terminology

- GradeClass: The classification of the grades of students based on GPA.
- Parental Support: The involvement and assistance provided by parents in the student's academic life.
- Extracurricular Activities: Non-academic engagements like sports, music, and volunteering, which may influence performance.

Costs and Benefits

- Costs:
  - Time spent cleaning and analyzing the dataset.
  - Computational resources for running models.
  - Stakeholder engagement to validate findings.
- Benefits:
  - Improved academic performance and reduced failure rates.
  - Increased parental awareness and involvement in education.
  - Enhanced decision-making for school administrators and policymakers.

Defining Your Data-Mining Goals

Data-Mining Goals

- Develop a predictive model that classifies students into GradeClass categories ('A', 'B', 'C', 'D', 'F') based on GPA.
- Find the most relevant predictors for academic performance, like StudyTimeWeekly and ParentalSupport.
- Offer actionable insights to better support students in the lower categories of performance.

Data-Mining Success Criteria

- The model will achieve at least 85% accuracy in predicting GradeClass on test data.
- Provide feature importance rankings to identify the top factors in performance.
- Validate model effectiveness through a simulation or case study using a subset of the data.

https://github.com/mattiasminejev/andmeteadus.git