Case Study for Prediction

This case study is designed to test your understanding and knowledge of data analytics, machine learning, and predictive modeling. The goal is to leverage your skills to generate actionable business insights and predictions to help business leaders make sound decisions.

Requirements

- Download and install Python: Python Download
- Install **TensorFlow** or **PyTorch**
- Install **Streamlit** for deployment

Project 1: Student Dropout Prediction

Problem Statement

Educational institutions need accurate predictive models to identify students at risk of dropping out. This project aims to develop and optimize predictive models to forecast student dropout rates and recommend interventions to improve retention and student success.

Objectives

1. Build an Accurate Predictor

- Develop machine learning models to predict student dropout.
- Aim for high accuracy in identifying students at risk of dropping out.

2. Optimize Model Performance

- Experiment with different model architectures to improve prediction accuracy.
- Fine-tune hyperparameters to enhance model performance.

3. Interpret Model Insights

- Create visualizations to understand and interpret model predictions.
- Gain actionable insights into factors contributing to student dropout.

4. Deploy the Model

- Develop a user-friendly interface for real-time dropout predictions.
- Ensure the deployed model is accessible and easy to use for educational institutions.

5. Generate Actionable Recommendations

- Based on model insights, recommend interventions to improve student retention.
- Provide strategies to support at-risk students and enhance overall student success.

6. Validate Hypotheses

- Test initial hypotheses about factors influencing dropout rates:

    a. Higher socio-economic status correlates with lower dropout rates.

    b. Students with higher admission grades are less likely to drop out.

    c. Dropout rates are lower among students receiving financial aid or scholarships.

7. Ensure Ethical Considerations

- Address potential biases in the data and model predictions.
- Ensure responsible use of predictive analytics in educational decision-making.

8. Develop a Comprehensive Solution

- Create a end-to-end pipeline from data preprocessing to model deployment.
- Provide documentation and guidelines for maintaining and updating the predictive system.

Hypotheses

1. Higher socio-economic status correlates with lower dropout rates.
2. Students with higher admission grades are less likely to drop out.
3. Dropout rates are lower among students receiving financial aid or scholarships.

Dataset

The datasets required for this case study are to be downloaded from 3signet.

Week 1: Data Collection and Data Wrangling

Tasks:

1. Environment Setup

    - Install Python, TensorFlow/PyTorch, and Streamlit
    - Set up a version control system (e.g., Git) for the project
    - Create a virtual environment for the project

2. Data Import and Initial Exploration

    - Download the dataset from 3signet
    - Load the data into a pandas DataFrame
    - Perform initial data exploration (shape, data types, summary statistics)

3. Data Cleaning and Validation

    - Handle missing values (imputation or deletion)
    - Identify and handle outliers
    - Correct data types (e.g., ensure dates are in datetime format)
    - Remove duplicate entries (if any)

4. Data Transformation

    - Normalize numerical features
    - Encode categorical variables (one-hot encoding or label encoding)
    - Create derived features (e.g., age from date of birth)

5. Statistical Analysis

    - Perform descriptive statistics on all variables
    - Conduct correlation analysis
    - Perform hypothesis tests (e.g., t-tests, chi-square tests) to validate initial hypotheses

Deliverables:

1. Jupyter notebook containing all data wrangling steps and code
2. Cleaned and transformed dataset (in CSV format)
3. Data preprocessing report (PDF) including:

    - Description of data cleaning steps
    - Summary of data quality issues encountered and how they were resolved
    - Justification for chosen data transformation methods

4. Statistical analysis report (PDF) including:

    - Descriptive statistics for all variables

- Correlation matrix heatmap
- Results and interpretation of hypothesis tests