Liina Hoogand
Kelli Gagarina

E4 report

Name:
Predict students' academic success
Link to repo:
https://github.com/liinahoogand/Data-science-project

Task 2. Business understanding

Identifying Our Business Goals

Background:

From the beginning we wanted to pick a topic for the project that would be interesting for us and also impactful, as life has shown that the more interested you are in a project - the better outcomes will arise. We decided to search for a dataset in Kaggle and the most relatable datasets were the ones that included university life, and so we stumbled upon a dataset with the name ”Predict students' dropout and academic success” and we’re really happy with it.

Business Goals:

We have set up multiple smaller goals for this project and dataset which all come down to the one big goal which is to learn from other students and see how we can make the best decisions for our academic lives. The smaller goals include understanding and analyzing the basics of our dataset, visualizing what kind of pool of people we are working with and then trying to find the biggest aspects that influence the dropout and graduate rate - for that we are planning to make dashboards, make a prediction model and making a hypothesis and testing it.

Business Success Criteria:

We have set the standards to ourselves that this project is successful when we have effectively visualized the dataset and gotten the necessary conclusions. But in the bigger perspective, we would consider this project successful when we have figured out the specific things that affect students dropout and graduation rate in an interesting manner.

Assessing Your Situation

Inventory of Resources:

Liina Hoogand
Kelli Gagarina

As mentioned before we have picked the dataset from Kaggle named ”Predict students' dropout and academic success” (link https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention) . This is the only dataset that we are going to look at for this project because we have many columns that we are very interested in analyzing and so are focused on this only. Due to the fact that there is a lot of work to be done to analyze all of these columns and find whether or not they are important, and if they are - then how so.

Requirements, assumptions, and constraints

Project Requirements:

Technological knowledge - Use appropriate tools and technologies for data analysis, dashboard creation, and prediction modeling.

Visualizations - Develop clear and informative visualizations, like dashboards, to show findings.

Collaboration - Use effective collaboration within the team to ensure the project gets done successfully.

Assumptions:

Data Accuracy - Assume that the dataset is accurate and representative of the factors influencing students' target.

Prediction Model - Assume that the constructed prediction model will provide valuable insights, although it may not guarantee absolute accuracy.

Constraints:

Time - Acknowledge that the project timeline is a constraint, requiring efficient task management and prioritization to meet deadlines.

Single Dataset Focus - Recognize that limiting the analysis to a single dataset may constrain the diversity of perspectives, requiring deep exploration of the available columns to compensate.

Risks and contingencies

1. Data Quality Issues

- Risk: The dataset may contain inaccuracies or missing values.

- Contingency: Implement thorough data cleaning processes, including validation checks.

Liina Hoogand
Kelli Gagarina

2. Modeling Complexity:

- Risk: The prediction model may not capture the full complexity of student behavior.

- Contingency: Continuously validate and refine the model.

3. Technology Failures:

- Risk: Technical issues may disrupt data analysis or model development.

- Contingency: Keep alternative tools or platforms in mind to quickly adapt in case of technology failures. Ask as soon as problems rise up and hold good communication between teammates.

Terminology

Dataset: A structured collection of data, in our case, the dataset named "Predict Students' Dropout and Academic Success" obtained from Kaggle.

Prediction Model: A statistical or machine learning model designed to predict outcomes, such as student dropout and academic success, based on historical data patterns.

Dashboards: Visual interfaces presenting key insights and trends derived from the dataset, facilitating easy interpretation of complex data.

Correlation: A statistical measure indicating the degree to which two variables change together, helping identify potential relationships within the dataset.

Hypothesis: A testable statement or assumption about the relationship between variables, forming the basis for further investigation and analysis.

Data Cleaning: The process of identifying and correcting errors or inconsistencies in the dataset, ensuring its accuracy and reliability.

Students target: The fact if the student has either dropped out, enrolled or graduated college. This is a term specific to our dataset.

Costs and benefits

Costs:

Time Investment: Extensive time will be required for data analysis, model development, and visualization creation.

Liina Hoogand
Kelli Gagarina

Training and Skill Development: Team members need to utilize all learned skills.

Benefits:

Informed Decision-Making: Insights derived from the analysis can inform academic institutions and students about factors influencing dropout and graduation rates.

Enhanced Understanding of Dataset: Thorough analysis will lead to a comprehensive understanding of the dataset and its implications.

Prediction Model Accuracy: A well-developed prediction model can provide valuable insights into potential student outcomes.

Dashboard Utilization: Visualizations and dashboards can facilitate easy communication of complex findings.

Defining your data-mining goals

Data-mining goals

1. Pattern Discovery: Uncover hidden patterns and trends within the dataset related to student behaviors, academic performance, and target.

2. Prediction Modeling: Develop a prediction model to forecast student targets.

3. Feature Importance Analysis: Identify the most influential features or variables in the dataset that significantly impact student retention and academic performance.

4. Validation and Model Evaluation: Ensure the reliability and generalizability of the developed prediction model.

Data-mining success criteria

Success criteria for the goals listed earlier:

1. Pattern Discovery: Identification of previously unknown patterns and trends in student behaviors and academic performance. Demonstration of the relevance and significance of discovered patterns to dropout and academic success indicators.

2. Prediction Modeling:Development of a prediction model with a high level of accuracy in forecasting student outcomes, including dropout and academic success.

3. Feature Importance Analysis: Identification of key features and variables that exert a substantial impact on student retention and academic performance.

4. Validation and Model Evaluation: Transparent documentation of validation methods and results, allowing for scrutiny and reproducibility

Liina Hoogand
Kelli Gagarina

Task 3. Data understanding

Outline data requirements
First step in gathering data is to outline data requirements. For our project we need data to consist of information about students(gender, nationality, age etc.), their status(dropout/graduate/enrolled) to create prediction models. Also social-economical and demographical data such as GDP, inflation rate and whether the student is debtor or not.
This would help to outline all contributors to potential academic success or failure. Time range could be at least 2 years to detect change in demographic data. Most suitable data format to work with would be a CSV.

Verify data availability
Dataset like this exists and is made available by platform Kaggle, which also provides links to the original source and has rights to share this dataset with everyone. It can be found on Kaggle under the name "Predict students' dropout and academic success". Only drawback is that this dataset contains information from 2 semesters - one academic year. Overall it doesn't much affect our goals, as we still can compare the impact of demographic aspects by comparing students from different regions.

Define selection criteria
We will use one(mentioned above) dataset, which consists of 4432 rows and 32 attributes. It is presented and can be downloaded in the .csv format. Fields of the dataset are: Marital status, Application mode, Application order, Course, Daytime/evening attendance, Previous qualification, Nationality, Mother's qualification, Father's qualification, Mother's occupation, Father's occupation, Displaced, Educational special needs, Debtor, Tuition fees up to date, Gender, Scholarship holder, Age at enrollment, International, Curricular units 1st sem (credited), Curricular units 1st sem (enrolled), Curricular units 1st sem (evaluations), Curricular units 1st sem (approved), Curricular units 1st sem (grade), Curricular units 1st sem (without evaluations), Curricular units 2nd sem (credited), Curricular units 2nd sem (enrolled), Curricular units 2nd sem (evaluations), Curricular units 2nd sem (approved), Curricular units 2nd sem (grade), Curricular units 2nd sem (without evaluations), Unemployment rate, Inflation rate, GDP, Target.

Describing data
Source of the dataset: https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention. At first, all attributes in the dataset will be used to detect specific ones that affect academic success. Most of the attributes are presented as numerical, even columns that hold non-numerical information such as Nationality, Marital status and Course. The description(decoding) for these attributes is available on Kaggle and will be also provided in the introduction to our project.

Exploring data
During data exploration we found out that the distribution of women is higher, most of the student's nationality is Portuguese as data was taken

from the universities in Portugal and as expected the majority of students were aged 18-23.
Dataset does not contain any Nan values. Distribution of the target is not balanced, as in the majority of cases student status is “Graduated” and might be balanced in order to create a more accurate prediction model.

Verifying data quality
Data quality is high, it combines information from multiple researches. The only minor quality issue that we noticed might be the variety of categorical values for attributes such as occupation and qualification of a student's parents. But this problem can be eliminated by feature engineering.

Liina Hoogand
Kelli Gagarina

Task 4. Planning your project

1. Writing an introduction Kelli 1h
2. Describing the data Kelli seaborn, matplolib 2h
3. Exploring main basic relations/correlations + visualizing Liina 6h
4. Creating a static dashboard Liina 4h
6. Creating interactive/dynamic dashboard Liina 6h
7. Preparing data for prediction model Kelli 3h
8. Compare prediction models sklearn for prediction models Kelli 4h
9. Describe result, finding best hyperparameters with randomized search Kelli 5h
10. Making a hypothesis and testing it Kelli 4h
11. Writing a conclusion Kelli Liina 2h each
12. Making/designing a poster 4h each
13. Preparing for poster session 2h each

ps1 Due to the fact that projects are often unpredictable and that we have to use a new unfamiliar software we have to take into consideration that there will be time being wasted just to "put out fires" and so I would consider 4h of extra time each.
ps2 There will also be time spent on just getting together and updating team members of progress and to discuss project goals and so I will add 5h for each.