PREDICTING FUTURE STUDENT (UAIRs) ENROLLMENT

Team Name : Data Magnates

Abstract

This project leverages machine learning techniques to predict student enrollment on UAIRs (University Analytics and Institutional Research) data set, aiding resource allocation, financial planning, and academic program development. By analyzing a comprehensive dataset containing academic, demographic, and socioeconomic factors, models such as Logistic Regression, Random Forest, and LSTMs were evaluated for accuracy and performance. Despite challenges with imbalanced data, the study demonstrates the potential of predictive analytics in supporting data-driven decision-making for institutional growth.

Team Members

Sahil More
Koumudi Vellanki
Karan Salot
Bavya Nalajala
Sharanya Neelam
Bhavana S

Table of Contents

1. Introduction & Motivation

Our project aims to predict future student enrollment at the University of Arizona, a critical endeavour for effective resource allocation, financial planning, and academic program development. Accurate enrollment forecasts enable the university to optimize class sizes, allocate faculty appropriately, plan campus facilities, and maintain financial stability. Ultimately, these insights will enhance the student experience and help ensure the university remains competitive in attracting and retaining students in an ever-evolving educational landscape.

To achieve this, we will apply a range of machine learning techniques, including Linear Regression, Decision Trees, Random Forests, Gradient Boosting, and Long Short-Term Memory (LSTM) networks. These models will allow us to capture both linear and non-linear patterns in the data, such as historical enrollment trends, economic factors, and demographic shifts. By leveraging these techniques, we aim to uncover the underlying drivers of enrollment trends and improve prediction accuracy.

We will conduct a series of experiments to compare the performance of these models, incorporating feature engineering, cross-validation, and time-based data splits to ensure robust and generalizable results. This approach will help us identify the most effective model for forecasting future enrollment, providing the University of Arizona with actionable insights for future academic and operational planning.

2. Dataset Description

For our project, we are utilizing data provided by the University of Arizona's Institutional Research and Analytics (UAIR) department. The dataset contains a rich set of attributes that are critical for accurately forecasting student enrollment. These attributes include information about students' academic careers, program details, enrollment status, and demographic data, which will allow us to analyse trends and patterns in the student population.

The key attributes in the dataset include Academic Career, which refers to the student's enrollment level (e.g., undergraduate, graduate); Academic Program Campus, which indicates the specific campus the student is enrolled in; and College, which reflects the academic college the student belongs to. Additionally, UA Full-Time Part-Time indicates whether the student is enrolled full-time or part-time, and Reporting Residency provides information on the student's residency status (in-state or out-of-state). The Headcount attribute represents the number of students enrolled, while Cohort refers to the group of students starting at the same time.

The dataset also includes demographic details such as IPEDS Race/Ethnicity Description to help analyse enrolment trends across different demographic groups, Term, which

indicates the academic term for enrollment, Gender to capture gender distribution, and Pell Eligibility Flag, which indicates whether a student is eligible for Pell Grants, providing insight into the socioeconomic status of students. This comprehensive dataset offers the necessary information to build a robust predictive model that can accurately forecast future enrollment trends based on various academic, demographic, and financial factors.

3. Data Preprocessing

To ensure the quality and accuracy of our machine learning models, we begin with a thorough data preprocessing process. First, we check for missing values in the dataset using the isnull() function, which identifies any missing or incomplete data across all columns. After identifying the missing values, we address them by filling missing categorical values with the placeholder "Unknown." This ensures that our models can handle missing data without issues that could arise from incomplete entries.

Next, we check for duplicate rows in the dataset using the duplicated() function. Duplicate records can lead to biases in the model training process, so we drop any duplicates using the drop_duplicates() function to ensure that each data point is unique and that our model is trained on accurate data.

The preprocessing steps also include encoding categorical variables and scaling numerical features. Categorical variables, such as "Gender," "Term," and "College," are converted into numerical values using label encoding, which transforms textual data into a format that can be used by machine learning algorithms. Numerical features like "Headcount" are scaled using StandardScaler to ensure that the model does not assign disproportionate weight to features with larger ranges. Finally, we split the dataset into training and testing sets using train_test_split(), with 80% of the data allocated to training and 20% reserved for testing. This ensures that we can evaluate the model's performance on unseen data, allowing for a more accurate assessment of its predictive capabilities.

4. Methodology

4.1 Model Building and Evaluation

In this project, we will apply various classification algorithms to predict "Full-Time vs. Part-Time Enrollment." The goal is to utilize a range of machine learning models to evaluate their performance and select the most suitable one for this task. The models we will apply include Logistic Regression, Support Vector Machines (SVM), Decision Trees, and k-Nearest Neighbors (k-NN). These models will be trained using the training dataset, and their performance will be evaluated using accuracy, precision, recall, and F1-score.

To optimize the models, we will apply cross-validation and hyperparameter tuning. Specifically, we will perform grid search on the k-NN model to find the best combination of hyperparameters. This helps to fine-tune the model and improve its predictive accuracy. After training, we will evaluate each model's performance on the test dataset by calculating the accuracy and generating classification reports.

The models applied are:

1. **Logistic Regression**: A linear model used to predict binary outcomes based on one or more predictor variables.
2. **Support Vector Machine (SVM)**: A powerful classifier that finds the optimal hyperplane to separate the classes with the largest margin.
3. **Decision Tree**: A tree-like structure that makes decisions based on input features, providing interpretability and simplicity.
4. **k-Nearest Neighbors (k-NN)**: A non-parametric model that classifies data based on the closest training examples in the feature space.

Each model's performance is evaluated based on accuracy and other classification metrics such as precision, recall, and F1-score.

4.2 Ensemble Methods

In addition to individual classifiers, we will apply ensemble methods to improve model performance. Specifically, we will use Random Forest and Bagging Classifier. Random Forest is an ensemble of decision trees, where each tree is trained on a random subset of the data, and their predictions are averaged to improve accuracy. Bagging Classifier also uses an ensemble of decision trees but with a focus on reducing variance by averaging the predictions across many trees.

Both models will be evaluated on the same test data, and their classification reports will be analysed to compare their performance to the individual models.

4.3 Clustering

We will explore clustering methods to identify patterns and groupings in the data. The clustering techniques used are K-Means, Hierarchical Clustering, and DBSCAN. Each of these algorithms has different strengths and weaknesses in terms of handling data structure, and we will evaluate their performance using the silhouette score, which measures how well-defined the clusters are. K-Means is a centroid-based method, Hierarchical Clustering builds a tree of clusters, and DBSCAN identifies clusters based on density, making it more adaptable to clusters of varying shapes and sizes.
The silhouette score will be used to assess how well each model groups the data, with higher values indicating better clustering.

4.4 Natural Language Processing (NLP)

If the dataset contains text-based features, we will perform basic Natural Language Processing (NLP) tasks. This may involve vectorizing the text data using techniques like CountVectorizer and TF-IDF Vectorizer. These methods convert text into numerical representations, which can be used for further analysis or modelling.

Additionally, we will demonstrate language translation using a pre-trained model, such as the transformers library, to showcase an application of NLP. For instance, we might translate English text into French to demonstrate multilingual capabilities of NLP models.

4.5 MLOps: Model Pipeline and Deployment

In the final step, we will implement an MLOps pipeline to automate the preprocessing, model training, and evaluation process. This pipeline will enable seamless integration of model building into the workflow. Additionally, we will demonstrate how to save and load models using joblib, making it easier to deploy the models in production environments. This ensures that once the models are trained, they can be reused and maintained efficiently. The entire pipeline will be integrated into the system to handle new data in the future, ensuring continuous improvements in model performance.

By following these steps, the project aims to apply machine learning to predict enrollment status while ensuring scalability, optimization, and robustness through model tuning, evaluation, and MLOps practices.

5. Experiments and Results

We tested multiple classification algorithms to predict "Full-Time vs. Part-Time Enrollment" based on student data.

1. **Logistic Regression**: Achieved 66% accuracy with a precision of 0.66 for full-time and 0.64 for part-time students. The recall for part-time students was low at 0.32, indicating difficulty in identifying part-time enrolments.
2. **Support Vector Machine (SVM)**: Also achieved 66% accuracy, with similar precision (0.66 for full-time, 0.65 for part-time). The recall for part-time students was 0.33, showing challenges in predicting part-time students.
3. **Decision Tree**: Had a lower accuracy of 51%, with a precision of 0.59 for full-time and 0.38 for part-time. The recall for part-time students was particularly low at 0.36.
4. **k-Nearest Neighbors (k-NN)**: Showed 60% accuracy, with better performance for part-time students (recall of 0.45). Hyperparameter tuning (neighbors = 9) slightly improved accuracy to 62%.

5.1 Ensemble Methods

1. **Random Forest**: Achieved 54% accuracy, with precision of 0.62 for full-time and 0.43 for part-time students. The recall for part-time students was 0.40.
2. **Bagging Classifier**: Similar to Random Forest with 54% accuracy and precision of 0.62 for full-time and 0.43 for part-time. The recall for part-time students was 0.41.

5.2 Clustering Analysis

1. **K-Means**: Generated poor clusters with a silhouette score of 0.14, indicating poorly separated clusters.
2. **Hierarchical Clustering**: Also yielded low separation (silhouette score of 0.16), suggesting difficulty in defining meaningful clusters.
3. **DBSCAN**: Performed poorly with a silhouette score of 0.05, indicating that DBSCAN did not form effective clusters.

The models struggled with the imbalance between full-time and part-time students, with Logistic Regression and SVM showing the highest accuracy but low recall for part-time students. Ensemble methods like Random Forest and Bagging provided more balanced predictions, but did not significantly outperform the individual classifiers. Clustering techniques also showed poor results, indicating the data might not naturally separate into meaningful clusters. Future work could involve further model optimization, class balancing techniques, or feature engineering to improve accuracy for part-time student predictions.

6. Discussion

The results of our experiments demonstrate the effectiveness of ensemble methods, such as Random Forest, in handling complex, non-linear relationships within the enrollment data. The inclusion of demographic and economic indicators, such as residency status and program type, significantly improved the models' predictive power, aligning with findings from previous research. While the clustering analyses provided useful insights into the structure of the data, further optimization of parameters, especially for DBSCAN, is necessary for better clustering results.

Several challenges arose during the project. Managing imbalanced data, such as the disproportionate number of full-time vs. part-time students, was a significant hurdle. Additionally, ensuring that the models generalized well to unseen data required careful attention to time-based validation, especially for long-term forecasting.

7. Conclusion and Future Work

This project illustrates the potential of machine learning in addressing key challenges in education, specifically predicting student enrollment. By accurately forecasting future trends, our models can provide the University of Arizona with valuable insights to support resource planning, program development, and institutional growth. Future work will focus on enhancing these predictions by incorporating external datasets, such as economic indicators, to improve forecasting accuracy. Additionally, the framework could be extended to predict program-level enrolments, allowing for more granular planning. In the long term, we aim to deploy these predictive models as part of a real-time decision-support system for university administrators, enabling them to make data-driven decisions on enrollment and resource management.

8. References

1. Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.
2. Shmueli, G., Patel, N. R., & Bruce, P. C. (2016). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. Wiley.
3. Witten, I. H., Frank, E., & Hall, M. A. (2011). "Data Mining: Practical Machine Learning Tools and Techniques."
4. University of Arizona Institutional Research and Analytics (for datasets)

9. Appendix

1. Code GitHub Link: https://github.com/karansalot/DataMagnates_MIS-545

2. Comprehensive Literature Survey- Theoretical Foundations

Data Mining, Machine Learning and Predictive Analytics

- Han, Kamber, & Pei (2011) - Foundational text on data mining techniques
- Fayyad et al. (1996) - Seminal work on knowledge discovery in databases
- Provost & Fawcett (2013) - Data-analytic approaches in institutional decision-making.

3. Detailed Methodology- Ensemble Learning Techniques and Big Data Analytics
   - Breiman (2001) - Pioneering work on Random Forest algorithms
   - Comprehensive exploration of ensemble method effectiveness in predictive modeling
   - Chen et al. (2012) - Comprehensive survey of big data challenges and opportunities
   - Exploration of large-scale data analysis in institutional contexts

4. Supplementary Results
    - Confusion Matrices: Confusion matrices for the best-performing models to provide deeper insights into model accuracy.
    - Clustering Visualizations: Silhouette score plots or cluster distribution graphs from clustering algorithms like K-Means, DBSCAN, and Hierarchical Clustering.

5. Data Information
    - Feature Engineering Insights: Strategies used for feature creation and selection. Highlight the impact of features like residency status, academic program, and Pell Grant eligibility on model performance.
    - Imbalanced Data Handling: Methods applied to address class imbalance, such as over-sampling, under-sampling, or using class weights.

6. Tools and Resources
    - Libraries and Frameworks: Scikit-learn, TensorFlow and visualization tools like Matplotlib, Seaborn
    - Data Sources: University of Arizona's Institutional Research and Analytics department.

Team Contributions:

Sahil More- Report and Presentation
Karan Salot - Coding and Presentation
Bhavana S- Survey and Report
Koumudi Vellanki- PPT and Illustrations
Sharanya Neelam- PPT and Additional Analysis
Bavya Nalajala - Implementation and Report Format