Final Report: Purdue Tuition

Authored by Benjamin Le

Problem statement

Purdue has upheld a tuition freeze for 14 consecutive years, a policy that encouraged accessibility and affordability. This strategy has strengthened Purdue's reputation and accessibility, but heavily limits their potential revenue growth and institutional expansion.

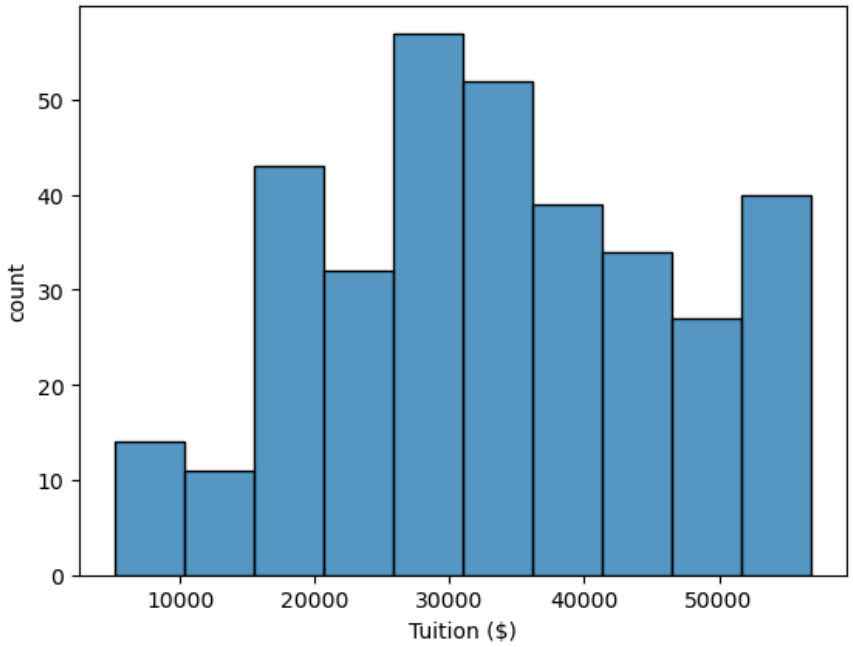
This report will investigate the opportunity cost of maintaining this policy, what potential avenues are recommended, and evaluate the key metrics in determining tuition pricing. The goal is to determine how Purdue can maximize its revenue through data-driven tuition pricing strategies and/or facility modifications, while maintaining accessibility for current and prospective students.

To accomplish this, the analysis will examine a range of variables that influence tuition pricing. The key question that will be explored is whether lifting or modifying the tuition freeze would ultimately yield positive net benefits for the university. By evaluating these factors, this report will provide a comprehensive understanding of the potential outcomes, risks, and advantages associated with either maintaining or breaking the tuition freeze.

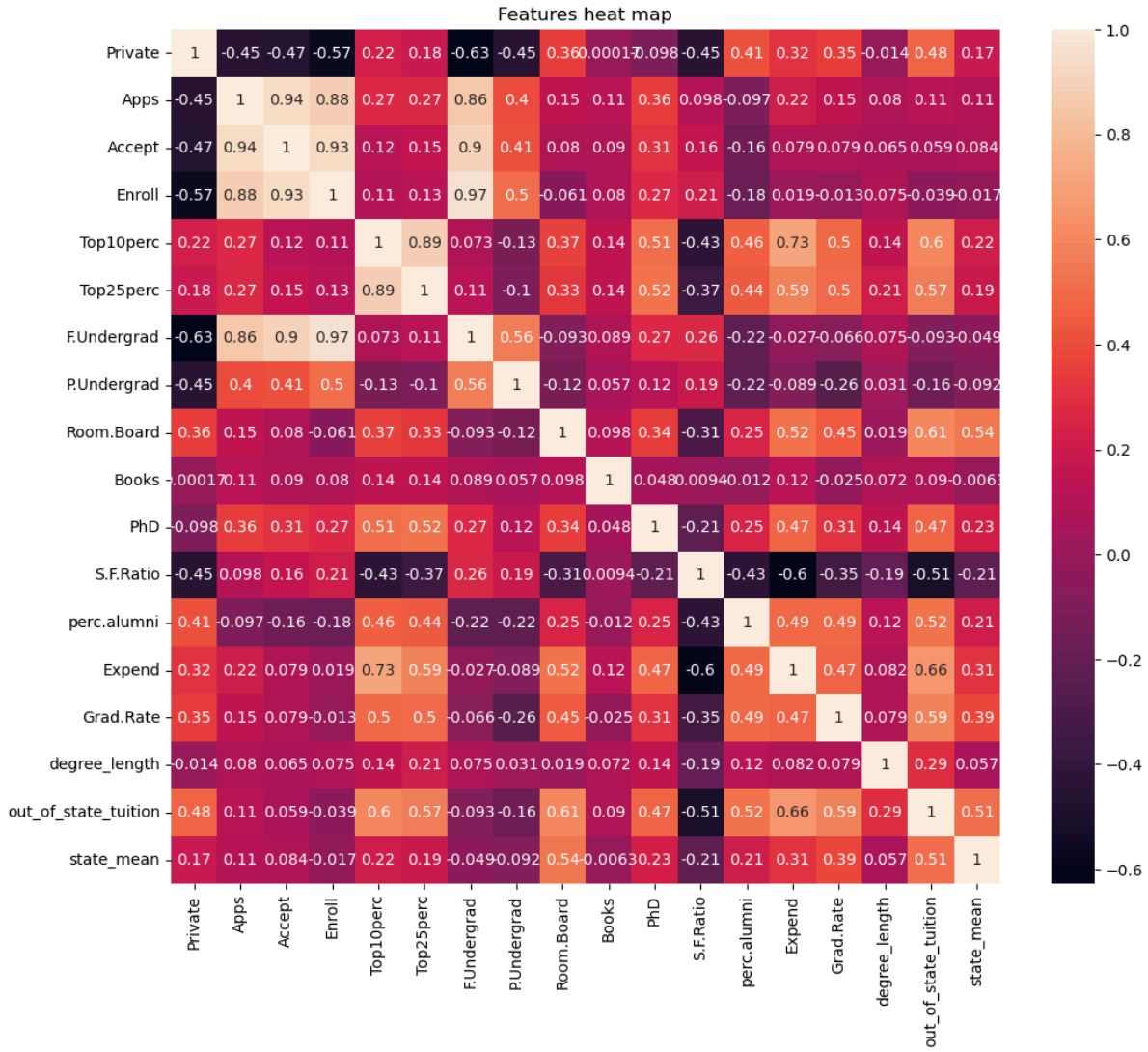
Data Wrangling

Two different datasets in 2018 were used for this analysis. The first has the core features of interest being evaluated (size: 777 x 16), and the second includes the tuition and state of the university (size: 2973 x 5). After merging and removing entries with missing or non-overlapping information, the final dataset consisted of **349 universities and 20 features**.

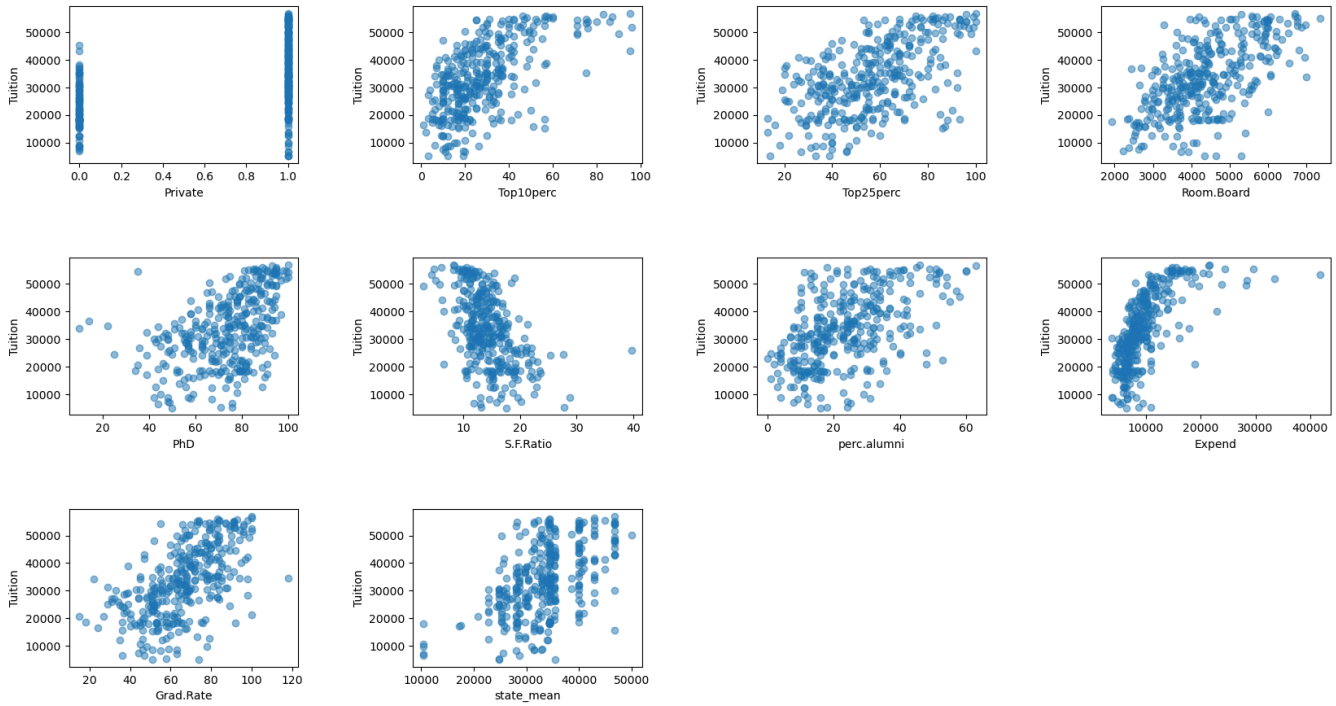
Exploratory Data Analysis



This displays the overall distribution of all the universities and their tuition. Purdue's tuition of $28,000 can be seen to lie on the lower end of the distribution, despite the university's strong academic reputation.



This heat map shows the full combination of correlations between all the features. When evaluating the feature of interest, out_of_state_tuition, it can be seen that the features 'Apps', 'Accept', 'Enroll', 'F.Undergrad', 'P.Undergrad', 'Books', and 'degree_length' are not significant features. These features were removed from consideration to avoid the curse of dimensionality.



These are the remaining individual correlations between the feature of interest and each significant feature. There is clear correlation between all features besides "Private" due to the nature of a binary variable.

The following is a full description of all of the features of interest:

Tuition: Out of state tuition

Private: private of public university (1 signifies private)

Top10perc: Percent new students from top 10% of H.S. class

Top25perc: Percent new students from top 25% of H.S. class

Room.Board: Room and board costs

PhD: Percent of faculty with Ph.D.'s

S.F.Ratio: Student/faculty ratio

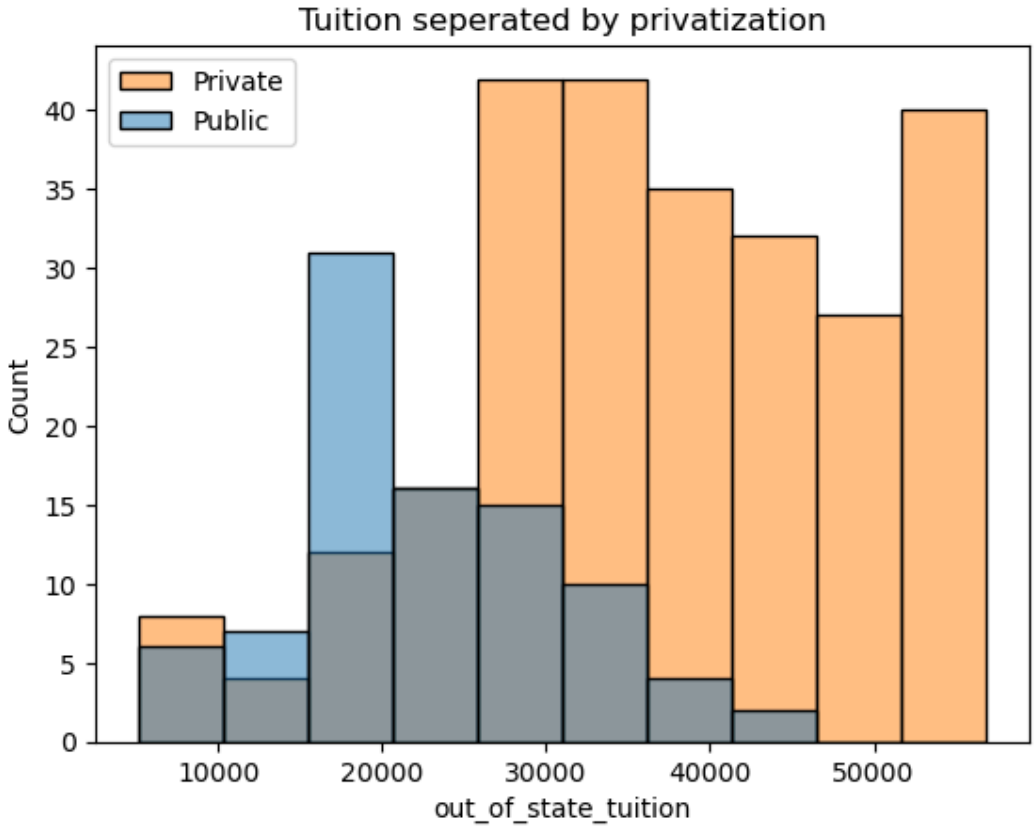
perc.alumni: Percent alumni who donate

Expend: Instructional expenditure per student

Grad.Rate: Graduation rate

state_mean: mean tuition of the respective state

A single incorrect graduation rate value (>100%) was manually corrected.



This is a better visualization of the “Private” feature.

PCA was explored, but using the PCA-derived features in the model was ultimately deemed unnecessary because the feature set of 10 was already manageable.

Modeling

The dataset was first split into training and test splits of 70/30 with the 'out_of_state_tuition' feature as the response variable.

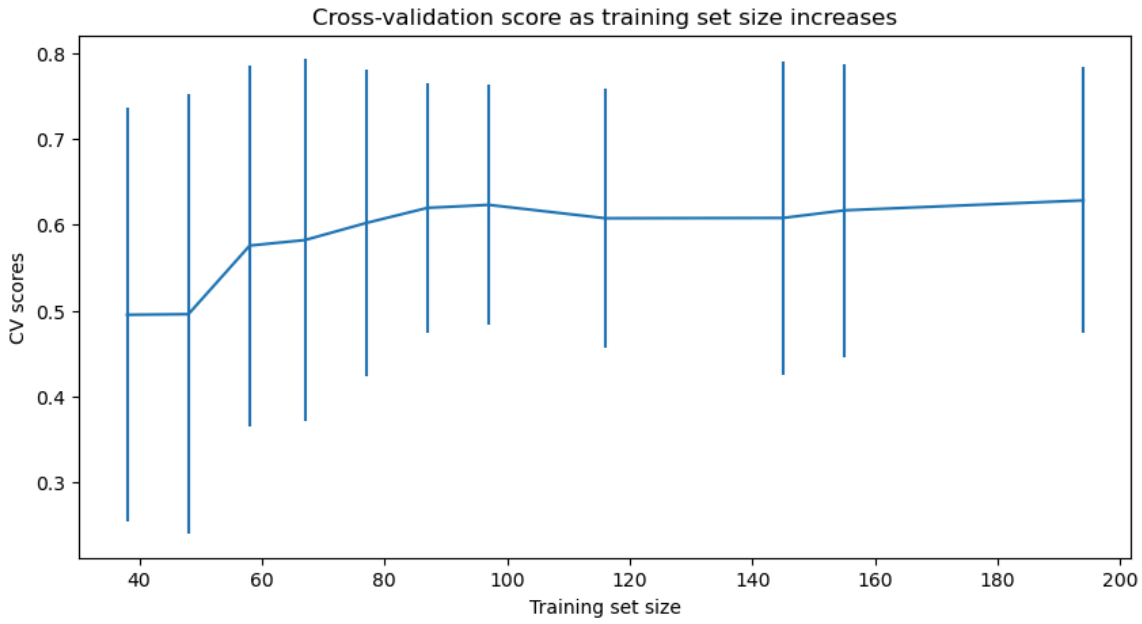
The two different models evaluated were:

1. Linear Regression model
2. Random Forest model.

GridSearchCV was used for hyperparameter tuning and cross-validation.

For the Linear Regression model, a pipeline of StandardScaler, SelectKBest, and the LinearRegression object was utilized. StandardScaler scaled the data and SelectKBest selected the k best features.

Similarly for Random Forest, a similar pipeline was made of SelectKBest and the RandomForest object. StandardScaler was not included as scaling is not necessary for a Random Forest model.



This graph is to test if there was enough data by testing how much the CV score changes as the training set size increases. It can be seen that there is not much change in CV score and its variance meaning there is sufficient data.

Findings

The Random Forest model performed the best with a cross-validated MAE of $5,230.81 (SD = 359.46).

The Linear Regression model had a comparable MAE of $5,375.50 (SD = 520.35).

The best hyperparameters for the Random Forest model found by the grid search were:

n_features: 10
max_depth: 10
min_samples_leaf: 1
min_samples_split: 5
n_estimators: 25

The model predicts Purdue's optimal tuition to be $33,712.03, compared to its actual tuition of $28,794.00. Even considering the model's mean absolute error of $4,472.43, Purdue is likely underpricing its tuition by over $5,000.

Recommendations

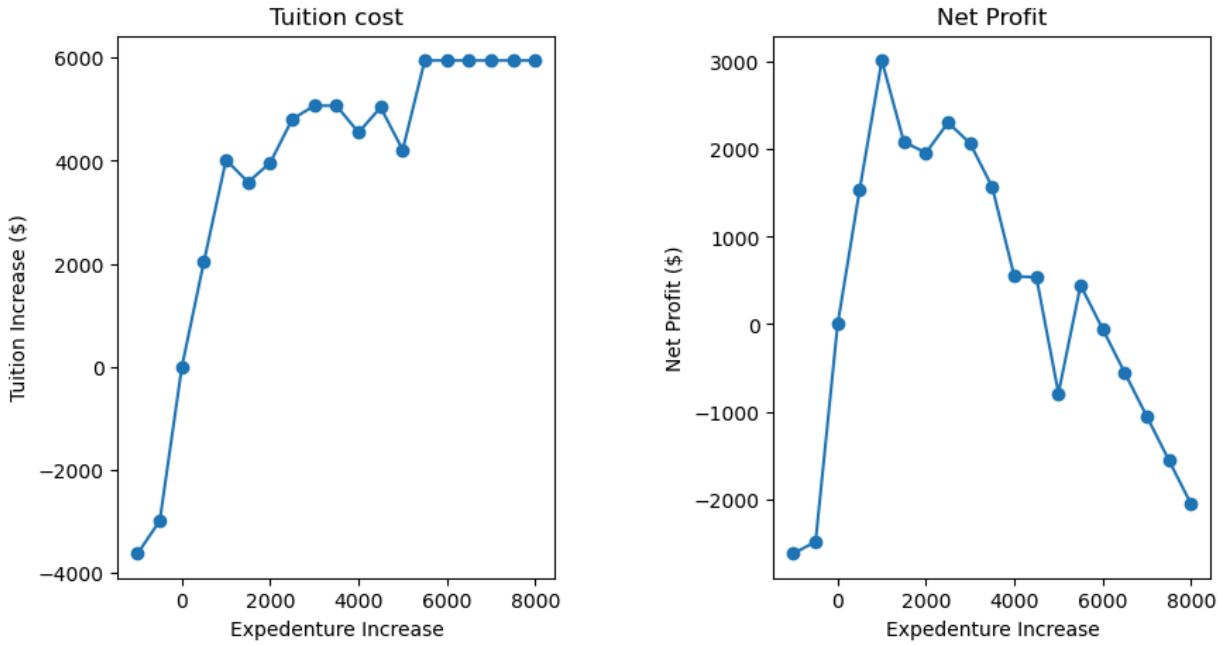
The central dilemma remains: Should Purdue maintain its tuition freeze at the cost of potential revenue, or adjust tuition to reflect its market value?

The model indicates a clear undervaluation of Purdue's tuition. While raising tuition involves reputational and accessibility concerns, the data suggests that Purdue can justify an increase without compromising institutional integrity.

Purdue can already justify tuition increases without major institutional changes. However, enhancing key factors that influence tuition could further strengthen this justification and potentially justify even higher tuition. The best model indicates that instructional expenditure and graduation rate are the primary drivers of tuition,

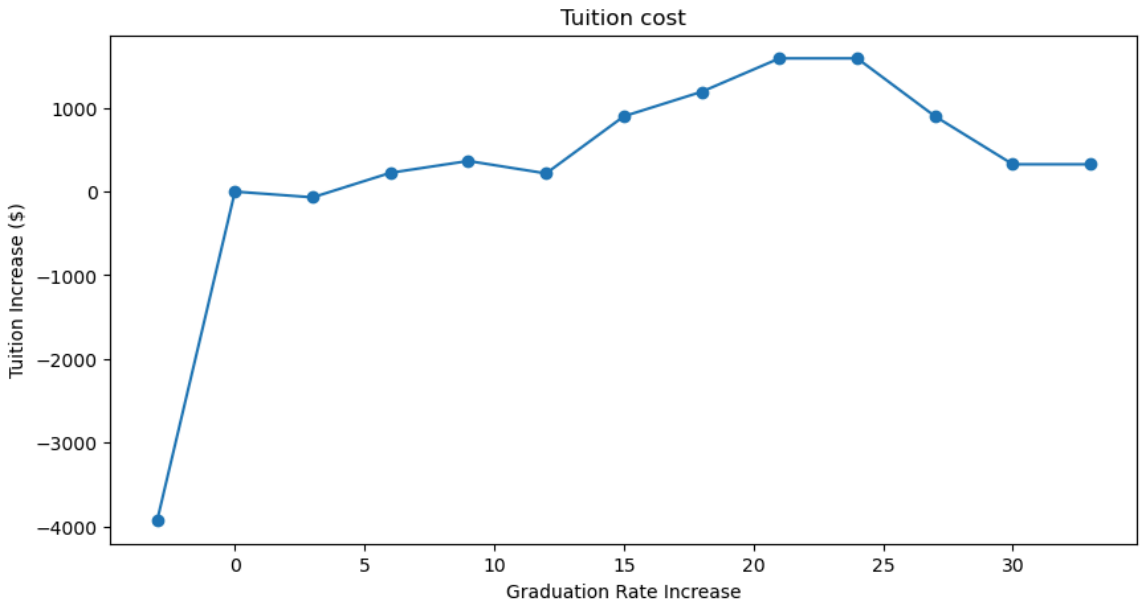
warranting further investigation. Currently, Purdue lies close to the center of both of these features' distributions suggesting room for improvement.

Regarding expenditure, the model recommends an increase of $1,000 in instructional expenditure per student which would justify a $4,000 increase in tuition.



These graphs show the tuition and net profit amount of different increments of expenditure increases. There is a clear plateau where an increase in expenditure spending will no longer influence tuition, but the elbow on the left graph shows the “best” possible decision.

Regarding graduation rate, the model recommends a 20% increase in graduation rate which would justify a $1,600 increase in tuition.



Similarly, to the expenditure recommendation, this graph shows the tuition value over different increments of graduation rate increases.

Further Research

Data Completeness: The model only evaluated 349 universities and many universities with missing data were removed. For a more comprehensive model, this missing data should be properly imputed with more universities included.

In-State Tuition: The model only evaluated out-of-state tuition; there could possibly be some nuance in dealing with in-state tuition although the out-of-state tuition is usually representative of the in-state tuition.

Revenue Impact: Estimating the actual financial return of tuition increases requires additional university-specific budget analysis.

Policy Implementation: How the university would actually apply improvements in expenditure spending and graduation rate is outside the scope of this study but should be investigated.