PREDICTING GRADUATION RATES

PROBLEM OVERVIEW

Academic success is a critical factor in bolstering socioeconomic growth. Significant dropout rates are a barrier that learning partners need to be empowered to address.

Earning Potential	College dropouts, on average, earn 32.6% than peers with an undergraduate-level education
Unemployment Rates	College dropouts are 19.6% more likely to be unemployed than any degree holder
Racial Disparity	Black students are 33.8% more likely to dropout than the average college student. American Indian/Alaska Native have a 45.1% dropout rate. White students are 7.9% less likely to dropout

DATA COLLECTION

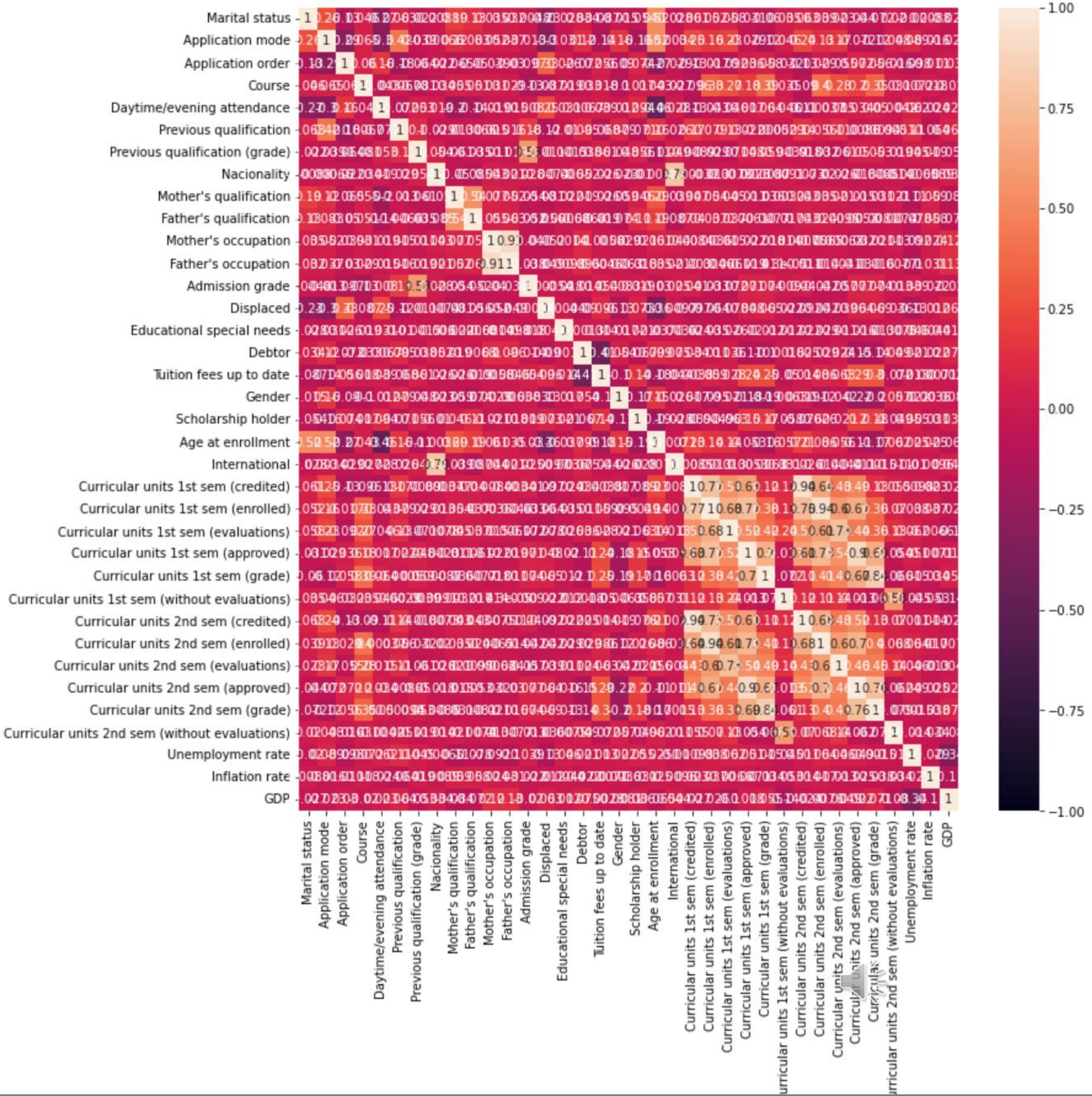
- Using collected data from UC Irvine, the dataset used to train the model holds 37 fields and over 5,000 rows of unique student data
- Includes socioeconomic data, enrollment data, and enrollment data at a period of time (1st and 2nd semester ends)

Table 1. Data Fields

Field Attribute	Field Name	Data Type
Demographic Data	Marital Status	Numeric
Demographic Data	Nationality	Numeric
Demographic Data	Displaced	Numeric
Demographic Data	Gender	Numeric
Demographic Data	Age at Enrollment	Numeric
Demographic Data	International	Numeric
Socieconomic Data	Mother's Qualification	Numeric
Socieconomic Data	Father's Qualification	Numeric
Socieconomic Data	Mother's Occupation	Numeric
Socieconomic Data	Father's Occupation	Numeric
Socieconomic Data	Educational Special Needs	Numeric
Socieconomic Data	Debtor	Numeric
Socieconomic Data	Tuition Fees Up to Date	Numeric
Socieconomic Data	Scholarship Holder	Numeric
Macroeconomic Data	Unemployment Rate	Numeric
Macroeconomic Data	Inflation Rate	Numeric
Macroeconomic Data	GDP	Numeric
Enrollment Data	Application Mode	Numeric
Enrollment Data	Application Order	Numeric
Enrollment Data	Course	Numeric
Enrollment Data	Daytime/Evening Attendance	Numeric
Enrollment Data	Previous Qualification	Numeric
Enrollment Data (End of 1st Semester)	Curricular Units 1st Sem (Credited)	Numeric
Enrollment Data (End of 1st Semester)	Curricular Units 1st Sem (Enrolled)	Numeric
Enrollment Data (End of 1st Semester)	Curricular Units 1st Sem (Evaluations)	Numeric
Enrollment Data (End of 1st Semester)	Curricular Units 1st Sem (Approved)	Numeric
Enrollment Data (End of 1st Semester)	Curricular Units 1st Sem (Grade)	Numeric
Enrollment Data (End of 1st Semester)	Curricular Units 1st Sem (Without Evaluations)	Numeric
Enrollment Data (End of 2nd Semester)	Curricular Units 1st Sem (Credited)	Numeric
Enrollment Data (End of 2nd Semester)	Curricular Units 1st Sem (Enrolled)	Numeric
Enrollment Data (End of 2nd Semester)	Curricular Units 1st Sem (Evaluations)	Numeric

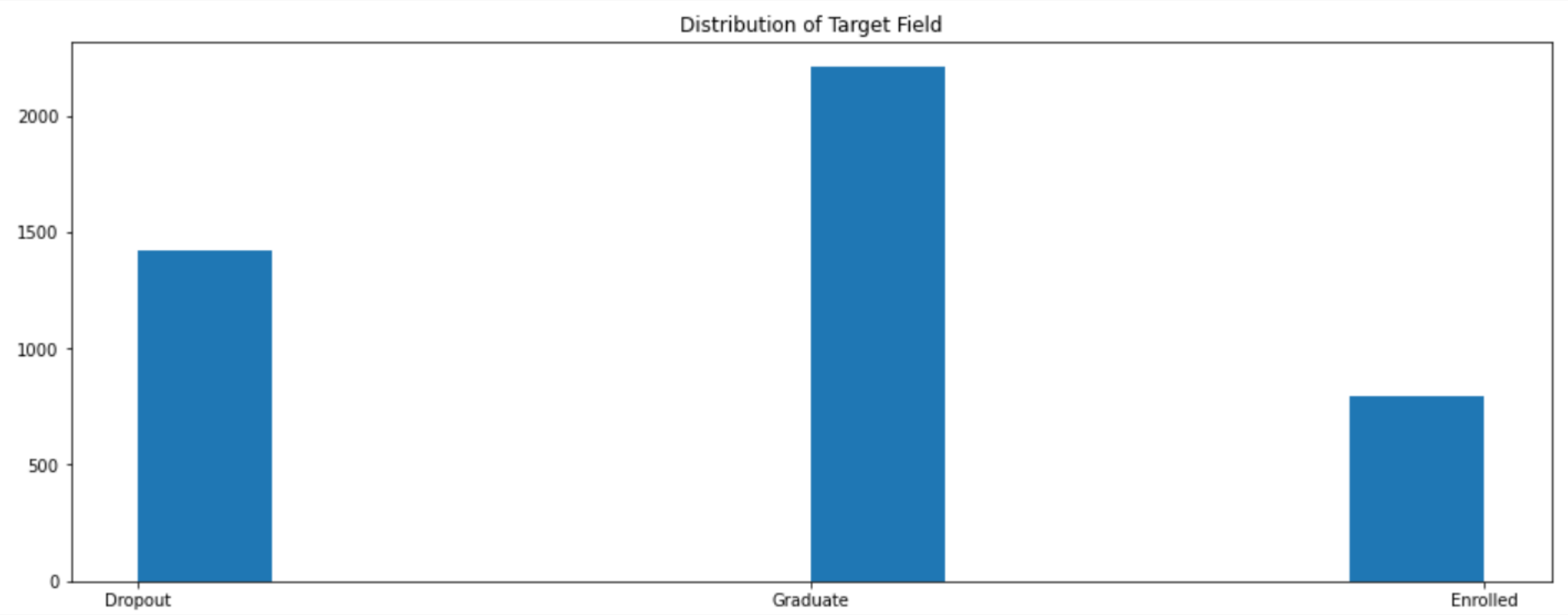
DATA CLEANSING AND PREPARATION

- Validated dataset for any NaN/Null values (no adjustments needed)
- Descriptive analysis was performed in order to better understand the basic statistics of the attributes (distribution, mean, median, min, and max)
- Evaluated the target field
- Constructed visualizations that address field correlation and multi-collinearity Early visualizations were created to look at field correlation and multi-collinearity



DEFINING THE TARGET

- Indicator of academic status- "Dropout", "Enrolled", and "Graduate"
  - Higher density of graduates (2,209) with dropouts next (1,421), and lastly enrolled (794). Graduates represented nearly 50% of the population, which could cause the model to skew towards a positive result.
  - Adjusted target field to binary (1 = "Graduate" or "Enrolled", 0 = "Dropout")



MODEL SELECTION + BUILD

Random Forest Classifier

- Ease of application
- Interpretability
- Higher level of accuracy than Decision-Tree Classifier alone

Base Model

```
rfc = RandomForestClassifier(random_state = 0) ## instantiate classifier
rfc.fit(X_train, y_train) ## fit model
y_pred = rfc.predict(X_test) ## predict Test set results
```

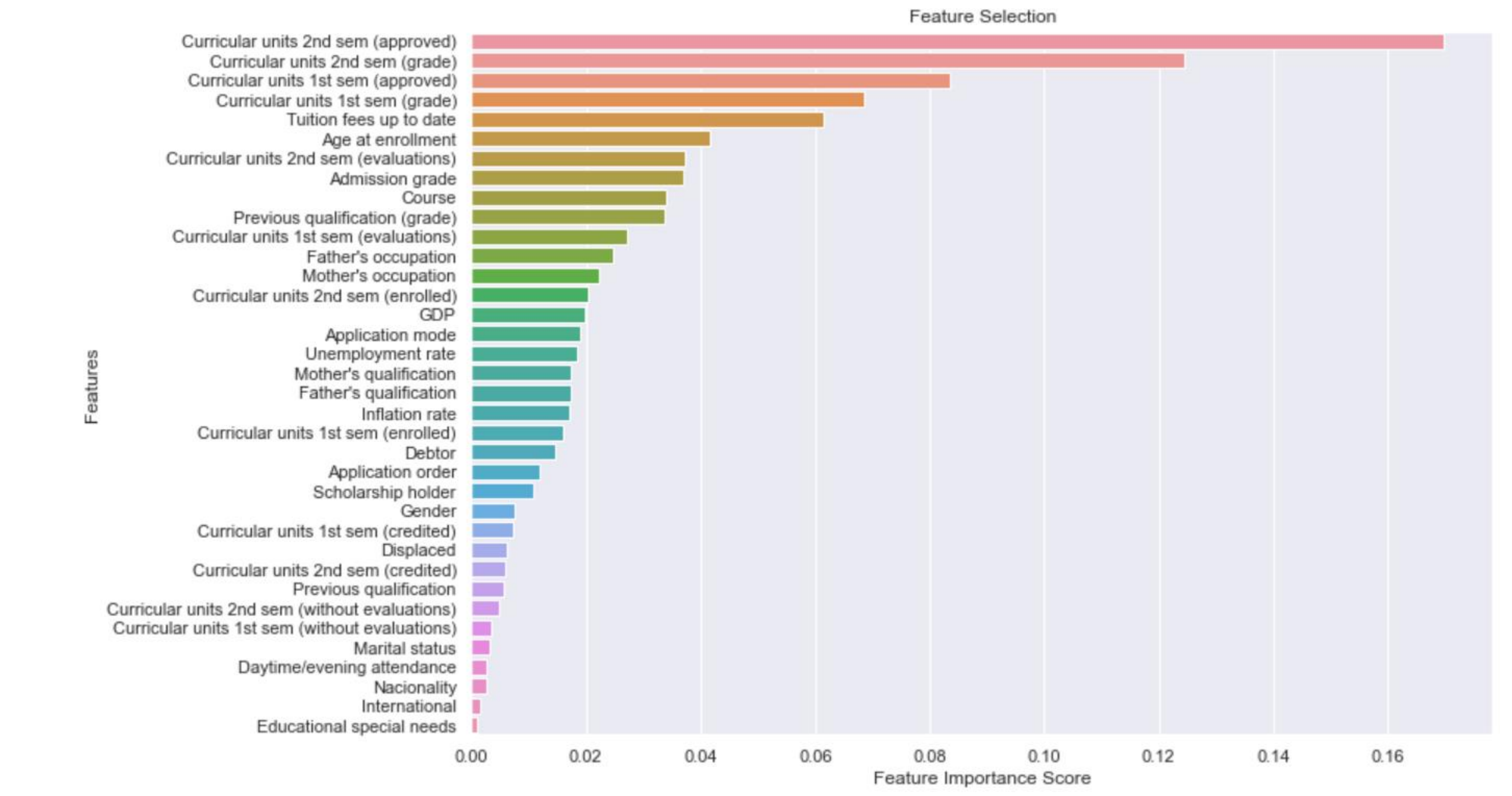
Initial Model Accuracy Score

- No quantifiable difference in adjusting the number of estimators

```
Model accuracy score with 10 decision-trees : 0.8554
```

```
Model accuracy score with 100 decision-trees : 0.8554
```

FEATURE SELECTION



```
Curricular units 2nd sem (approved)    0.169821
Curricular units 2nd sem (grade)       0.124344
Curricular units 1st sem (approved)    0.083448
Curricular units 1st sem (grade)       0.068416
Tuition fees up to date                0.061234
                                          ...
Marital status                         0.003309
Daytime/evening attendance             0.002624
Nacionality                            0.002515
International                          0.001437
Educational special needs              0.000895
```

ADJUSTING THE DATA

- Dropped 5 Lowest Fields
- 'Marital status’
- 'Daytime/evening attendance’
- ‘Nationality’
- 'International’
- 'Educational special needs’

Resulted in **~8% model accuracy improvement**

EVALUATING THE REBUILD

Confusion Matrix

228	88
34	535

- True Positive (Upper Left): 228, which indicates where the model correctly predicted the positive class
- False Positive (Upper Right): 88, which indicates where the model incorrectly predicted the positive class when it was actually negative (type 1 error)
- False Negative (Lower Left): 34, which indicates where the model incorrectly predicted the negative class when it was actually positive (type 2 error)
- True Negative (Lower Right): 535, which indicates where the model correctly predicted a negative class.

FURTHER FEATURE SELECTION

- Dropped an additional 10 columns where the importance of the field was less than the mean importance
  - 'Course'
  - 'Previous qualifications (grade)'
  - 'Admission grade'
  - 'Tuition fees up to date'
  - 'Age at enrollment'
  - 'Curricular units 1st sem (approved)'
  - 'Curricular units 1st sem (grade)'
  - 'Curricular units 2nd sem (evaluation)'
  - 'Curricular units 2nd sem (approved)'
  - 'Curricular units 2nd sem (grade)'

Resulted in **~10% model accuracy decrease**

```
## import SelectFromModel
from sklearn.feature_selection import SelectFromModel

sel = SelectFromModel(RandomForestClassifier(n_estimators = 100))
sel.fit(X_train, y_train)
```

```
SelectFromModel(estimator=RandomForestClassifier())
```

```
## True = importance is greater than the mean importance
## False = importance is lss than the mean importance
sel.get_support()
```

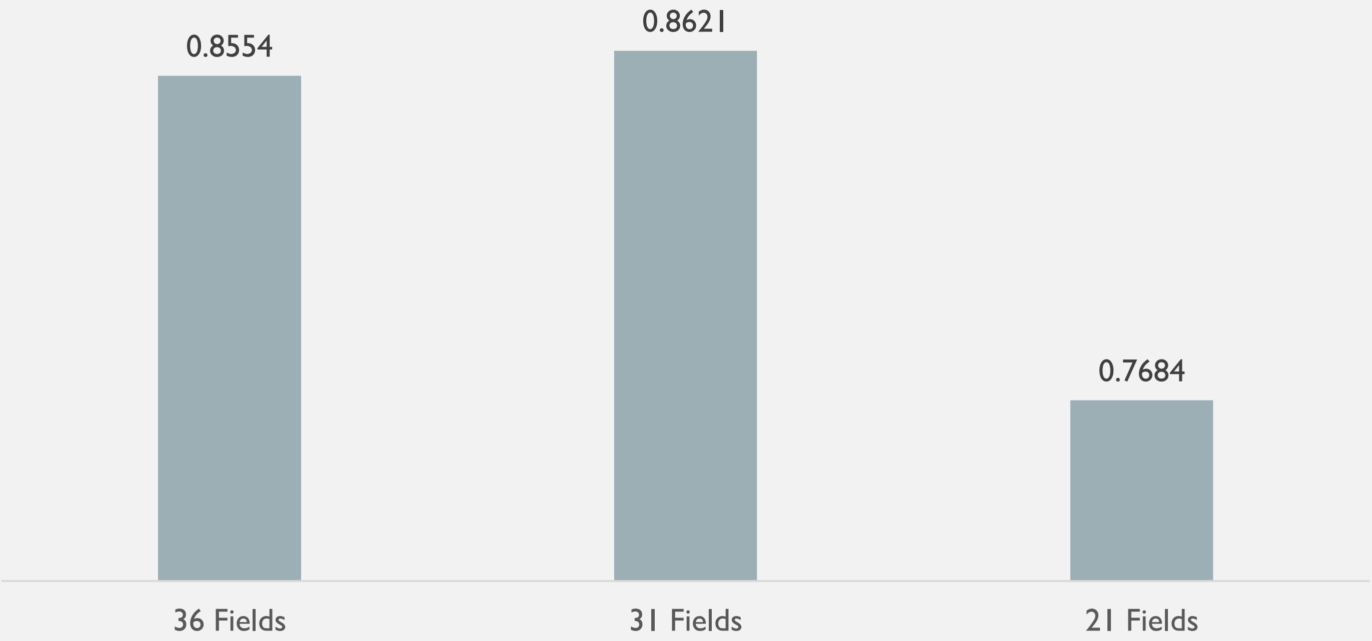
```
array([False, False,  True, False,  True, False, False, False, False,
        True, False, False,  True, False, False,  True, False, False,
       False,  True,  True, False, False, False,  True,  True,  True,
       False, False, False, False])
```

```
## get features column names
selected_feat= X_train.columns[(sel.get_support())]
len(selected_feat)
print(selected_feat)
```

```
Index(['Course', 'Previous qualification (grade)', 'Admission grade',
       'Tuition fees up to date', 'Age at enrollment',
       'Curricular units 1st sem (approved)',
       'Curricular units 1st sem (grade)',
       'Curricular units 2nd sem (evaluations)',
       'Curricular units 2nd sem (approved)',
       'Curricular units 2nd sem (grade)'],
      dtype='object')
```

COMPARING MODEL RESULTS

.8554 36 FIELDS	**.8621** 31 FIELDS	**.7684** 21 FIELDS



- 2nd model performed the best
- Over-reduction of fields resulted in a significant loss of collinearity

ETHICAL IMPLICATIONS AND LIMITATIONS

- Data is sensitive in nature
  - Project data was publicly sourced and without PII to protect student anonymity
- Ethical data collection and analysis pertinent to prevent bias, manipulation, or influence
- Educational ethics and considerations were the guidepost of analysis
  - Includes data caveats, limitations, and providing the audience with proper contextual information needed to navigate the nuances of any statements or conclusions
- Project understands that there are limits to how well the data can portray people and their actions, only meant to guide and inform learning providers
- Assuming that being a 'Dropout' is the antithesis to success is an oversimplification of the educational system done so to conduct this analysis and research

POTENTIAL FUTURE WORK

Expand Sample Dataset

- Project dataset was collected in partnership with UC Irvine and is limited to degree-seeking students
- Future workstreams could expand to different types of learning providers—public vs private, at the county, state, or other geographic determination, or by field of study

Partner with Learning Providers

- Partnership with learning providers / academic institutions could provide further insights into the student body and behaviors
- Further work could include expanding the potential field attributes or looking at populations of students that are more at-risk for dropping out

Different Models

- As a comparison, a decision tree model was built only resulted in a .8305 accuracy score, which was less than the accuracy score of the original model
- Future work could evolve this project's model's accuracy score or delve into alternative model options explored in earlier milestones

REFERENCES

- Hanson, M. (2022, June 17). College Dropout Rates. Retrieved from Education Data Initiative: https://educationdata.org/college-dropout-rates

- Hore, A. (2022, June). Predict Dropout or Academic Success. Retrieved from Kaggle: https://www.kaggle.com/datasets/ankanhore545/dropout-or-academic-success?select=Dropout_Academic+Success+-+Sheet1.csv

- National Forum on Education Statistics. (2010, February). The Forum Guide to Ethics. Retrieved from https://nces.ed.gov/pubs2010/2010801.pdf