Analysis of College Scorecard

Sheryan Resutov, Eugene Sokolov, Harrison Zhao

Dataset

- Kaggle dataset for 7804 U.S. universities
- Around 1700 features
- Data covers 1996 – 2013
- Features cover
  - About the school
  - Academics
  - Admissions
  - Costs
  - Student body
  - Financial aid
  - Completion
  - Earnings
  - Repayment

Problems With The Dataset

- Many elements are available only for students who receive federal grants and loans
- Treasury elements are protected for privacy purposes and shown as PrivacySuppressed
- Most elements are data pooled over two years to reduce year-over-year variability
- Many very specific features such as family income for 75k – 110k range

Transforming The Dataset

- Binning maps continuous values such as salary into discrete ranges
- $f_{max} - f_{min}$ are divided into n discrete bins
  - f_{max} is the maximum value for the feature
  - f_{min} is the minimum value for the feature
  - n is the number of bins
- For example, 43 binned into the range 40 – 50 with 5 bins would give a bin number of 2

Preliminary Algorithms Used

- Logistic Regression
- Support Vector Machines (SVM)
  - Grid search to find optimal parameters
- Five fold cross validation
- Both algorithms did not perform well for classification

Lasso Regression

- Least Absolute Shrinkage and Selection Operator
- Regularized least squares

$$\min_w ||Xw - y||_2{}^2 + \alpha ||w||_2{}^2$$

$$\min_w \frac{1}{2n_{samples}} ||Xw - y||_2^2 + \alpha ||w||_1$$

Random Forests

- How it works
  1. Bagging
  2. Random feature selection
  3. Ensemble learning
- Used 100 or 200 trees

Analysis Performed

- Analysis for:
  - Mean earnings
  - Loan default rate based on income bracket
  - Loan default rate based on major
  - Completion rate

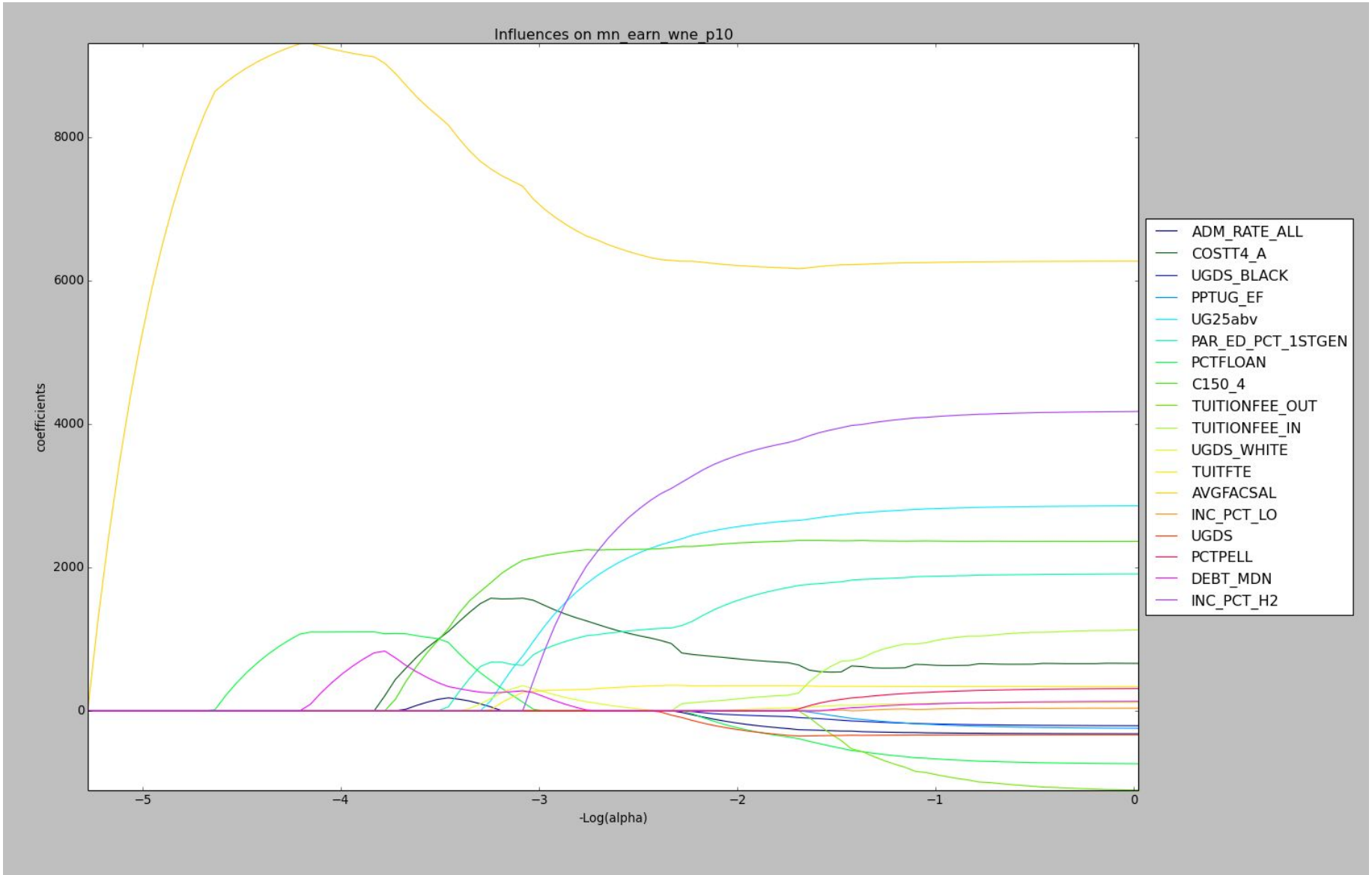
Mean Earnings Analysis

- Dependent Variable: Mean earnings of students working and not enrolled 10 years after entry (mn_earn_wne_p10)
- Independent Variables:
  - ADM_RATE_ALL (Admission Rate)
  - COSTT4_A (Average cost of Attendance)
  - UGDS_BLACK (% black students)
  - PPTUG_EF (% students part time)
  - UG25abv (% students age>25)
  - PAR_ED_PCT_1STGEN (% first gen)
  - PCTFLOAN (% receiving fed loan)
  - TUITIONFEE_OUT (Out-state tuition)
  - TUITIONFEE_IN (In-state tuition)
  - UGDS_WHITE (% white students)
  - TUITFTE (net tuition)
  - AVGFACSAL (faculty salary)
  - INC_PCT_LO (% poor students)
  - UGDS (total undergrad students)
  - PCTPELL (% receive Pell Grant)
  - DEBT_MDN (loan principal)
  - INC_PCT_H2 (% wealthy students)

Mean Earnings Analysis

- Using 8 bins
  - 13.4k bin size
  - 20.7k min
  - 128.4k max
- Classification accuracy using
  - Logistic Regression: 62%
  - SVM: 59.7%
  - Random Forest: 72%

Lasso Regression on Mean Earnings



Lasso Regression on Mean Earnings

- Used only TUITFTE and INC_PCT_H2
  - TUITFTE - Net tuition revenue per full-time equivalent student
  - INC_PCT_H2 - Dependent students with family incomes between $110,001+ in nominal dollars
- Accuracy with Random Forest increases to 78%
- The rich get richer

Loan Default Rate Analysis

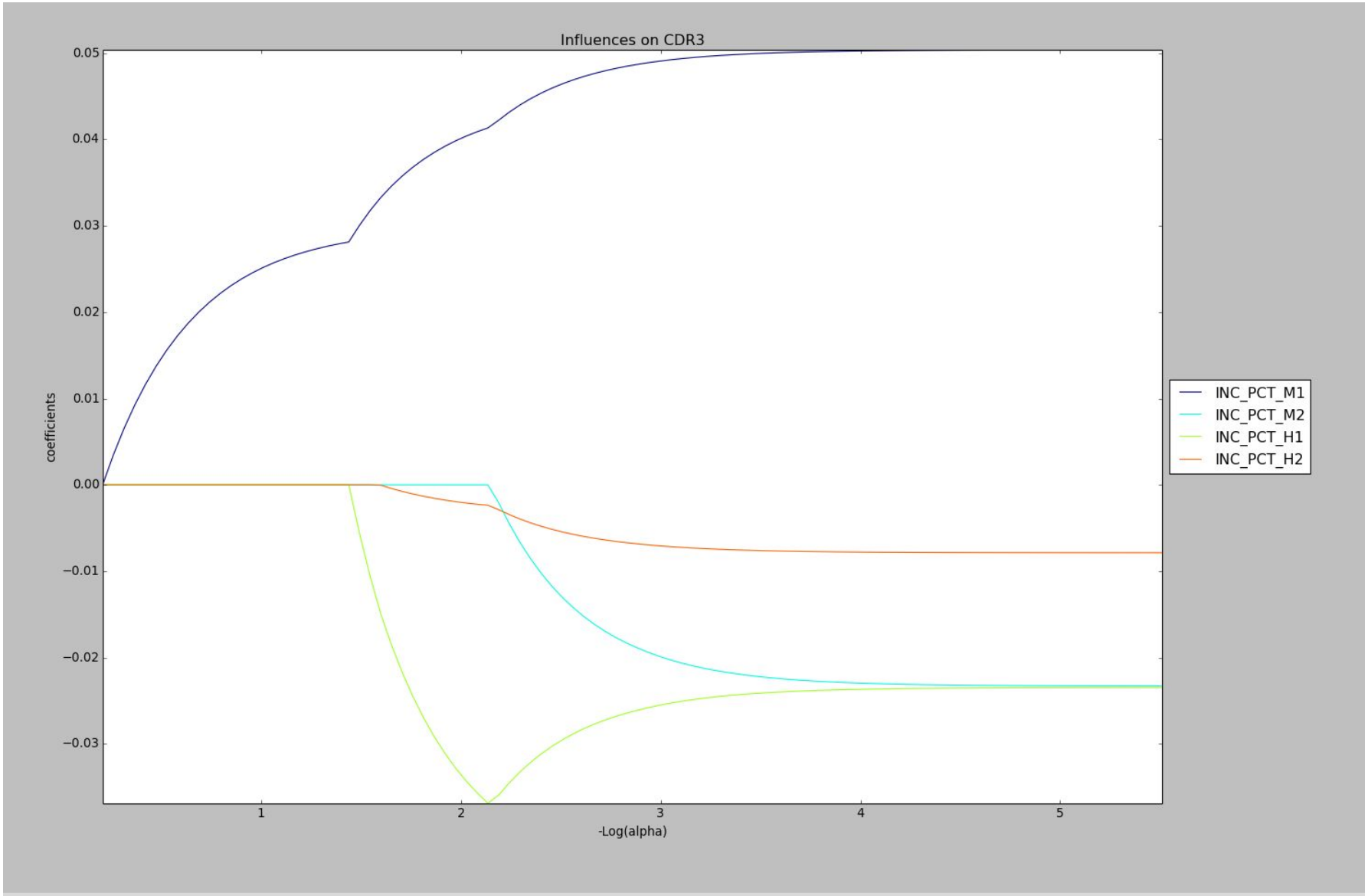
- Dependent Variable: Three-year cohort default rate (CDR3)
- Independent Variables:
  - INC_PCT_M1 *(between $30,001-$48,000)
  - INC_PCT_M2 *(between $48,001-$75,000)
  - INC_PCT_H1 *(between $75,001-$110,000)
  - INC_PCT_H2 *(between $110,001+)

*Dollar amounts correspond to aided students' family incomes

Loan Default Rate Analysis

- Using 10 bins
  - 0.2 bin size
  - 0 min
  - 1 max
- Classification accuracy using
  - Logistic Regression: 53%
  - SVM: 50%
  - Random Forest: 75%

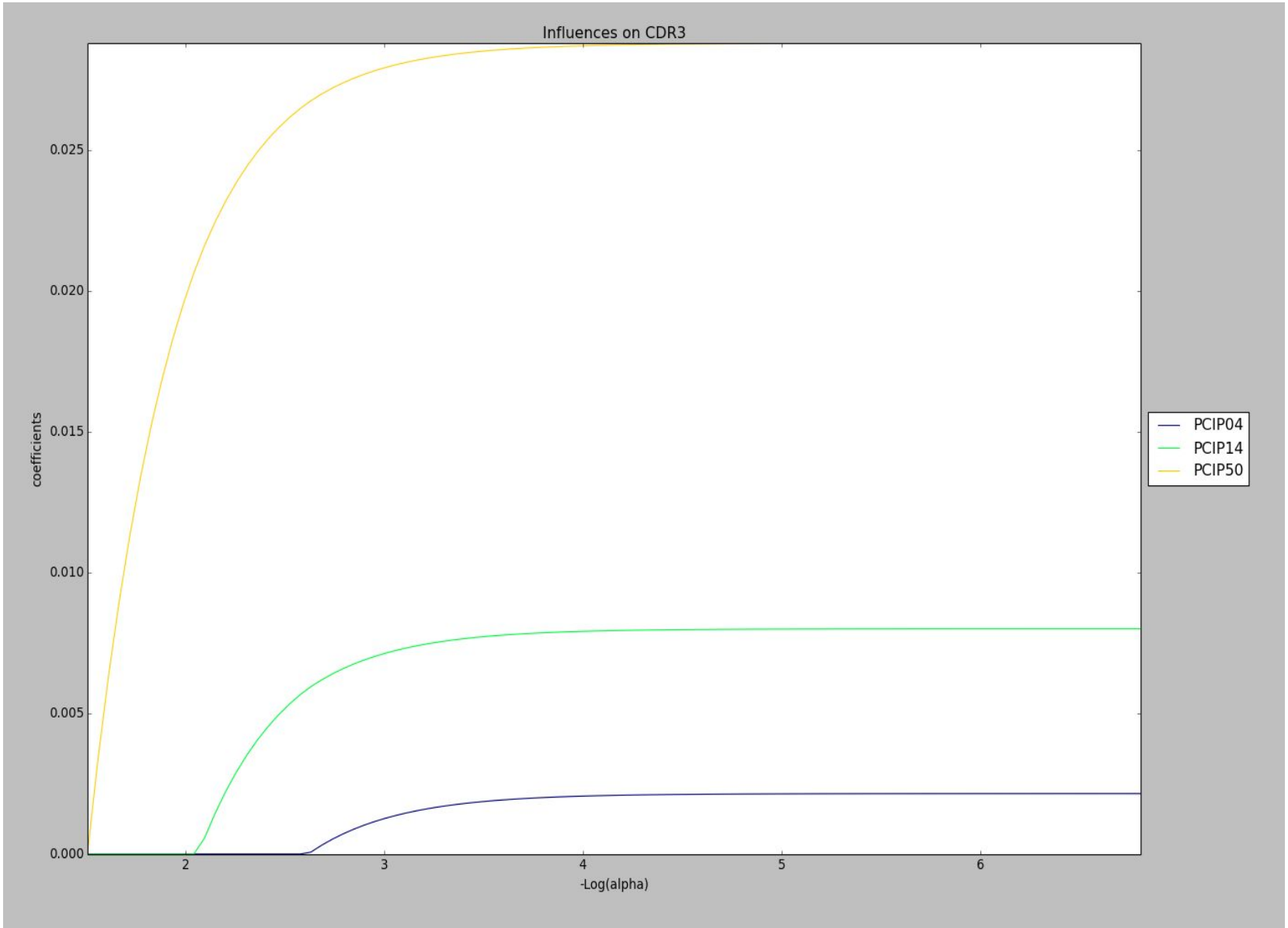
Loan Default Rate Analysis



Loan Default Rate Analysis

- Using only INC_PCT_M1 (aided family incomes $30,001-$48,000) accuracy goes down to 55% for Random Forest
- Perform Lasso regression based on the majors:
  - PCIP04 (% degrees Architecture)
  - PCIP14 (% degrees Engineering)
  - PCIP50 (% degrees Art)

Loan Default Rate By Major



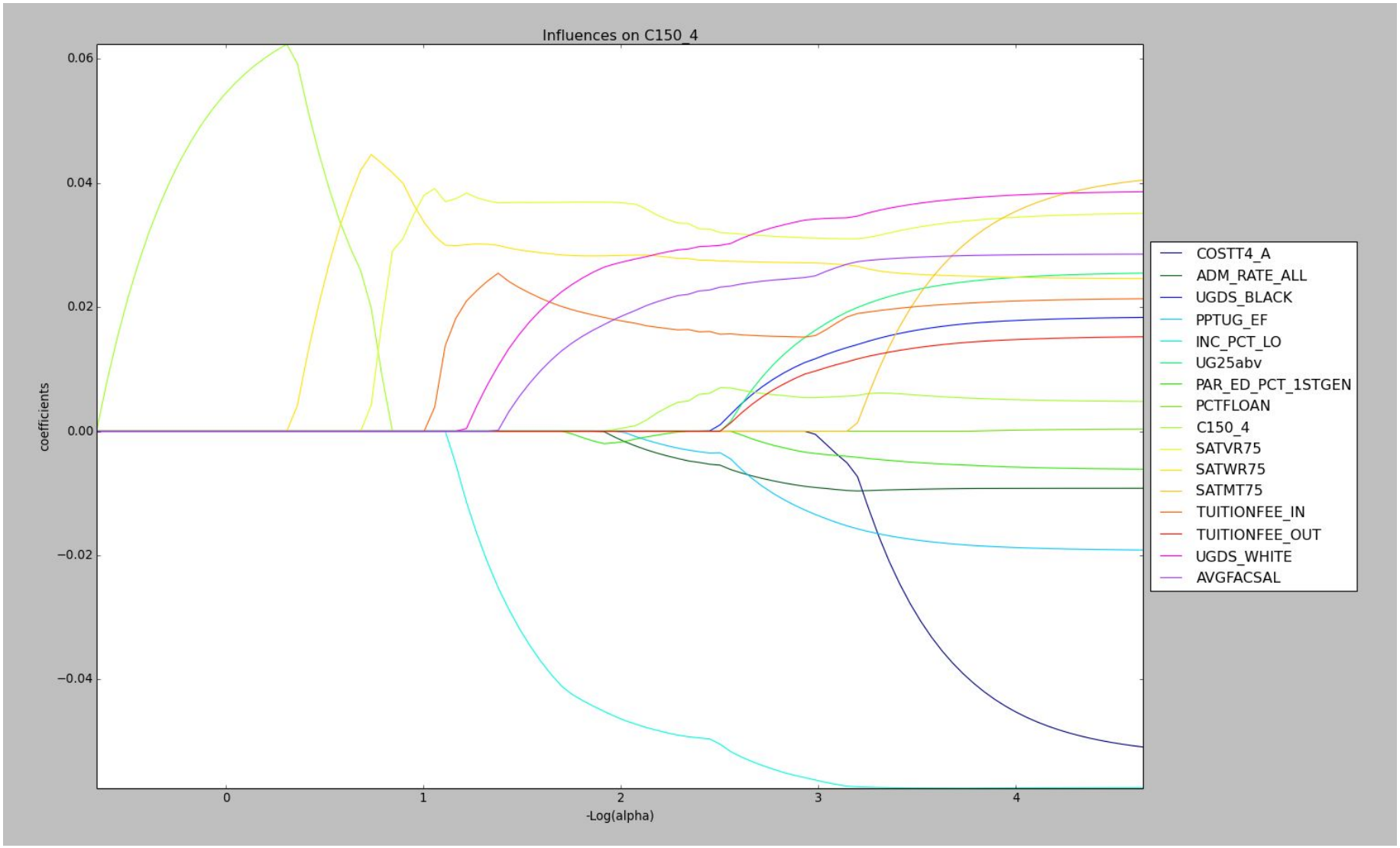
Completion Rate Analysis

- Dependent Variable: Completion rate for first-time, full-time students at four-year institutions (C150_4)
- Independent Variables:
  - COSTT4_A (Average cost of Attendance)
  - ADM_RATE_ALL (Admissions Rate)
  - UGDS_BLACK (% black students)
  - PPTUG_EF (% students part time)
  - INC_PCT_LO (% poor students)
  - UG25abv (% students age>25)
  - PAR_ED_PCT_1STGEN (% first gen)
  - PCTFLOAN (% receiving fed loan)
  - SATVR75 (Reading SAT 75th %tile)
  - SATWR75 (Writing SAT 75th %tile)
  - SATMT75 (Math SAT 75th %tile)
  - TUITIONFEE_IN (In-state tuition)
  - TUITIONFEE_OUT (Out-state tuition)
  - UGDS_WHITE (% white students)
  - AVGFACSAL (Faculty Salary)
  - TUITFTE (Tuition Fees)

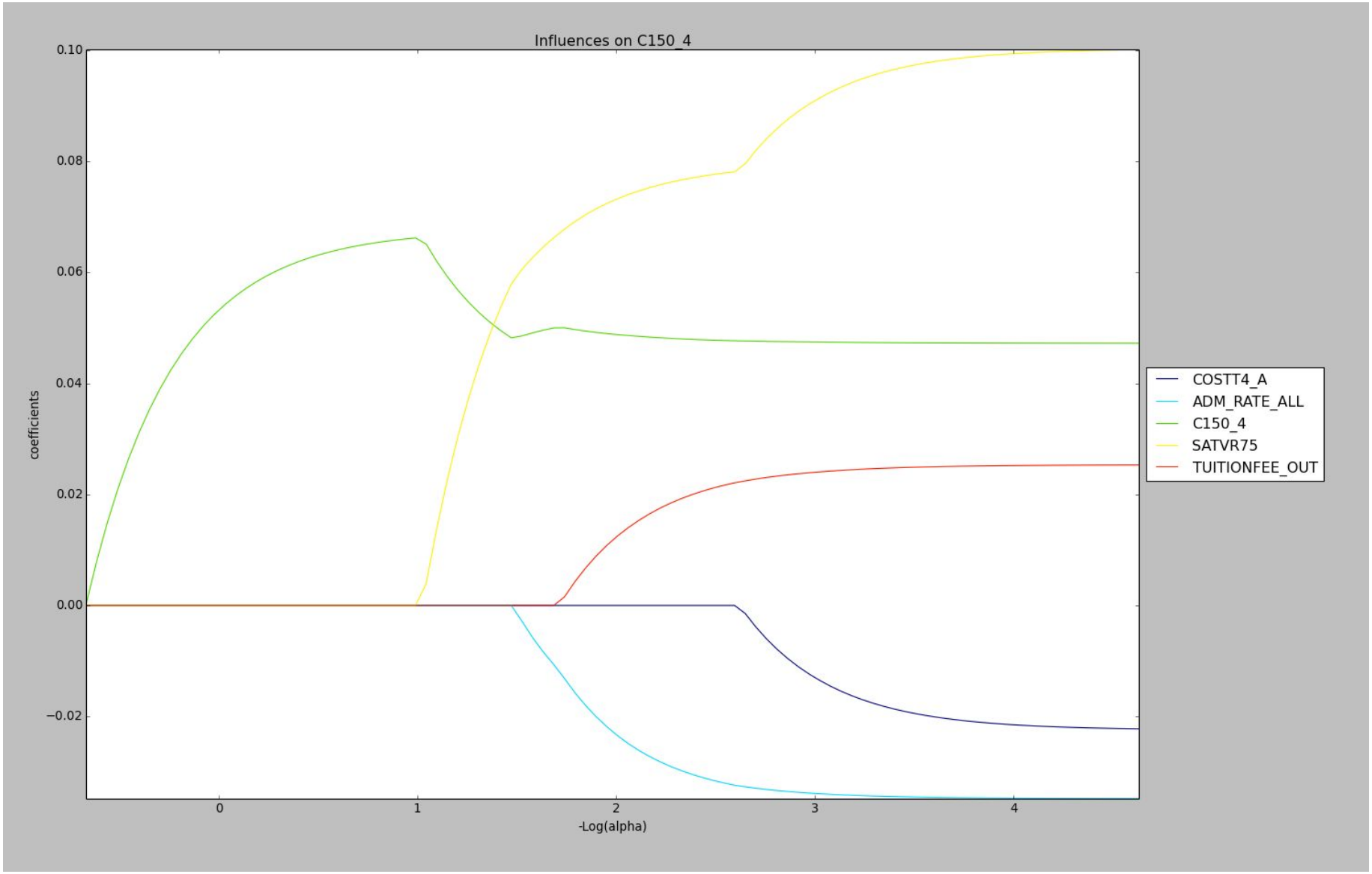
Completion Rate Analysis

- Using 5 bins
  - 0.18 bin size
  - 0.057 min
  - 0.9628 max
- Classification accuracy using
  - Logistic Regression: 45%
  - SVM: 37%
  - Random Forest: 62%

Completion Rate Analysis



Completion Rate Analysis



Completion Rate Analysis

- Lasso analysis independent variables:
  - COSTT4_A (Average cost of Attendance)
  - ADM_RATE_ALL (Admission Rate)
  - SATVR75 (Verbal SAT 75th percentile)
  - TUITIONFEE_OUT (Out-state tuition)
  - UGDS_WHITE (% white students)
- Classification accuracy using
  - Logistic Regression: 41%
  - SVM: 38%
  - Random Forest: 62%

Conclusion and Future Work

- Feature reduction
- Different classifiers
- Collecting more data
- Privacy suppression lowers classification accuracy

Thank You!