Predicting Student Churn From the University of Washington

Nishant Velagapudi, Lovenoor Aulck
Joshua Blumenstock, Jevin West

DataLab

Introduction

Nationally, 30% of undergraduate students do not return for a second year of college. Over $9 is spent in educational expenses on these students. Retention rate is a significant concern due to rising tuition costs and the growing necessity of earning a degree. Universities invest significantly in retention efforts: but identifying at risk students can be difficult. Here, we use data from the UW Registrar's office in a supervised Machine Learning approach to predict which students will churn (or "drop out") from the University of Washington. We aim to make our predictions based on data acquired after only a single quarter of being a student. We aim to identify at risk students early, when Universities will still be able to make impactful retention efforts.

Methods

- Prepare Data
  - Sample equal amounts completion/non-completion for a balanced data set
  - Filter out Tacoma & Bothell students: restrict data to 1998-2016 enrollees
  - Categorize data (first quarter GPA, first quarter classes, demographics, high school performance)
- Run parameter sweeps using KFCV
- Establish initial machine learning models using intuitive features
  - Logistic Regression
  - K-Nearest Neighbors
  - Random Forest
- Evaluate performance of each model

Results

Demographic Features:

	African American	American Indian	Asian	Caucasian	Race Not Known	Resident	Non Resident
Completion	72.80%	71.40%	81.40%	79.70%	79.90%	80.50%	74%
Count	2011	938	16037	39442	10242	61290	7828

Table 1: Lists of demographic variables in dataset: splits of graduation vs nongraduation and number of students in each category – note that race and residency are not exclusive

First Quarter Performance:

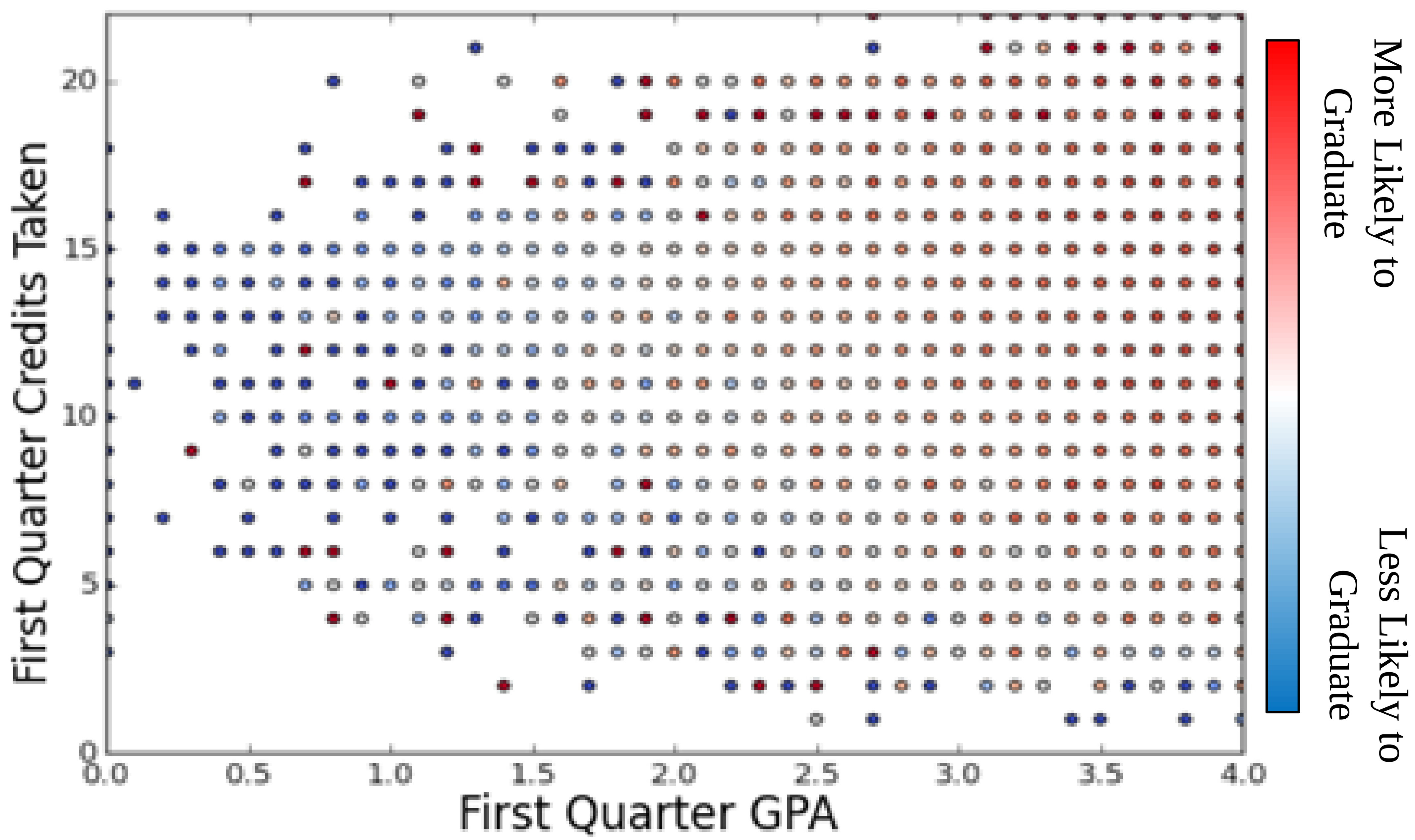

Figure 1: Color map showing distribution of completion at various GPAs earned and credits taken. Note that higher GPAs and more credits taken graduate more often.

Initial Machine Learning Models (On balanced dataset)

- Parameter Sweeps (10 fold cross validation):
  - LogReg: .01 regularization parameter (100x overfit)
  - K-Nearest Neighbors: 36 neighbors
  - Random Forests: 36 Trees

Model Performance (70-30 train test split)

- LogReg: 66.59% accuracy (.729 AUC)
- K-Nearest Neighbors: 64.60% accuracy (.660 AUC)
- Random Forests: 62.24% accuracy (.694 AUC)

ROC Curves:

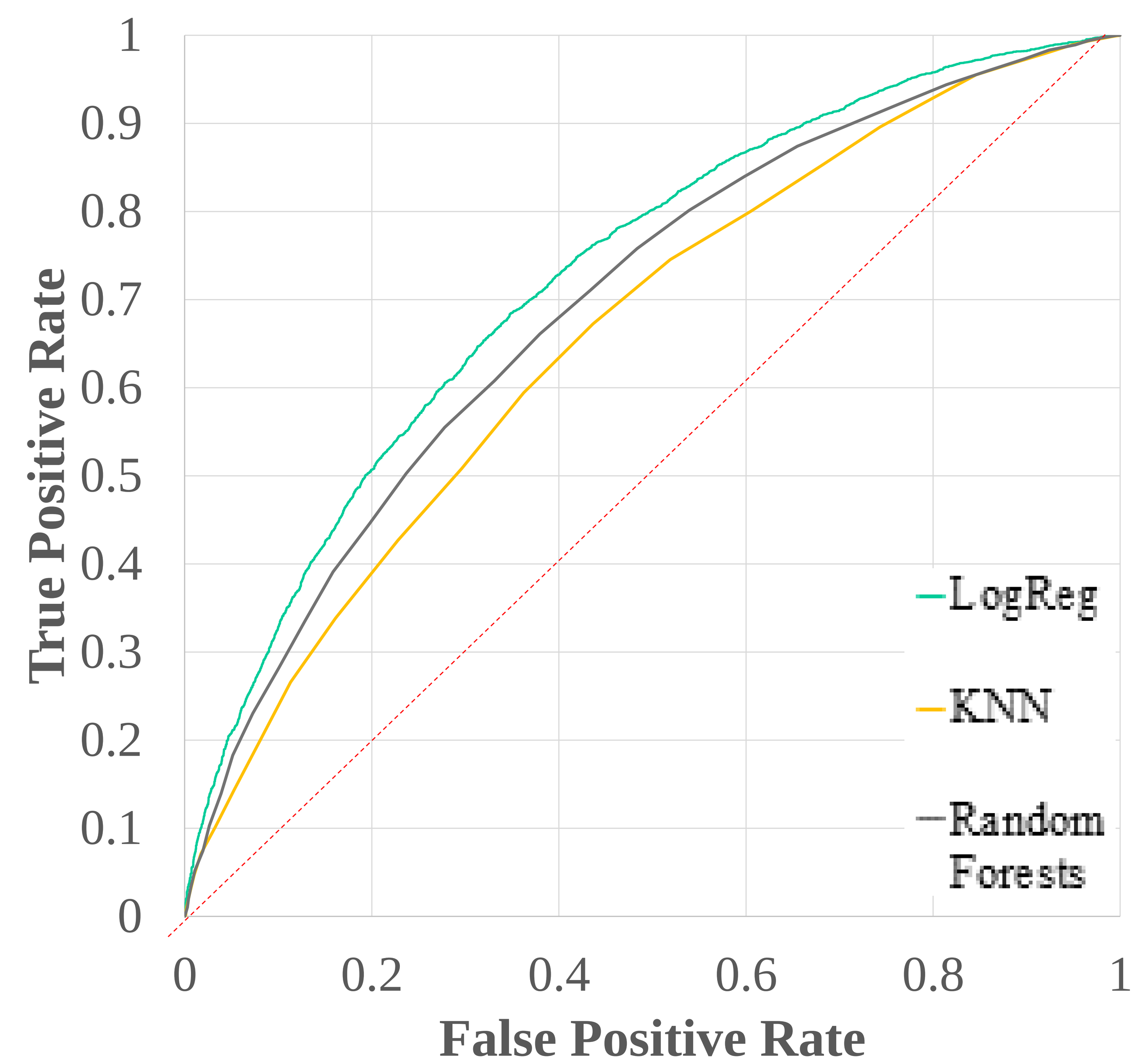

Figure 2: Performance in terms of Operating Characteristic Curves (ROC). We can see that LogReg is the most accurate algorithm.

Predicting Length of Stay

- Linear Regression to predict when dropouts leave school
- RMSE of ~5 quarters

Conclusions

- 16% boost over baseline accuracy
- Concerns of heterogeneity of student population
- Data does not cover important facets of student life
- Logistic Regression is the best performing algorithm

Acknowledgements: We would like to thank the UW registrar's office for providing anonymized student data for the purpose of these experiments