



UNIVERSITY *of* WASHINGTON

# LIMEaid

## Local Interpretable Model-agnostic Explanations (LIME)

Data 515, Spring 2019  
M.S. Data Science

Suman Bhagavathula  
Patrick King  
Javier Salido

# Interpretability in Machine Learning

Some highly accurate models are not “explainable”

- Neural networks, random forests

Why is this a problem?

- Bias, not obvious
- High test set accuracy but poor results in the field
- Policy or law demands an explanation of any decision

Solution: model-agnostic local explanations

- Explain one instance, not entire model
- Fit a simple model to explain a small section of decision space

# LIMEaid: A LIME solution for tabular data

## Input

- A “complex” ML model, fit by sklearn classifier object with .predict
- An instance of data ( $x$ ) and its model output ( $f(x)$ )
- Probability domain for normalized predictor variables (histograms)

## Output

- Sparse linear models (few features), plottable
- List significant features

## Analysis/verification

- Comparison to decision trees

# Use Cases

User profile: data scientist with Python programming experience

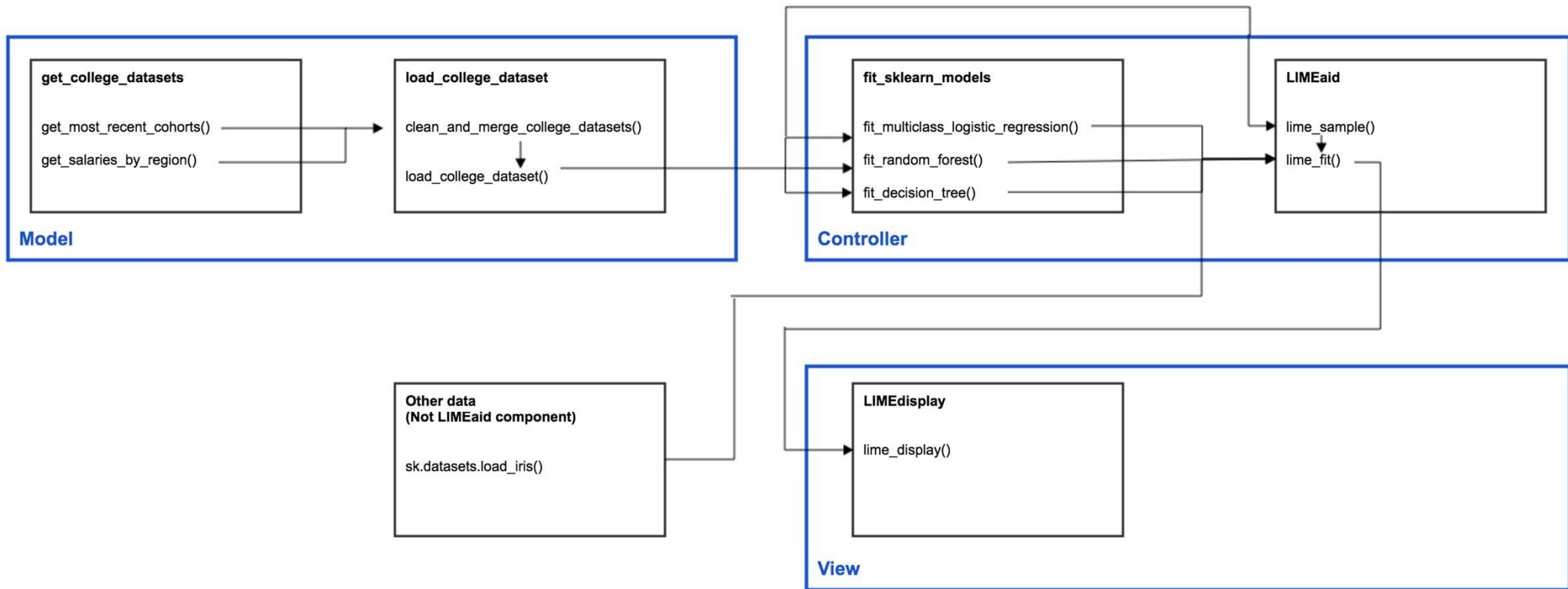
## 1. Model verification scenario

- User wants to preempt poor model performance “in the field”
- Use LIMEaid to sample test dataset
- Show most significant features for decision
- Tune or replace model if spurious correlation or other issues

## 2. Decision explanation scenario

- Classification has already been made by a model
- Use LIMEaid to sample whole dataset
- Produce easy-to-explain “two-dimensional” plot of linear regression

# Design



# Data

## Sources

- College Scorecard ([data.gov](https://data.gov)): Annual report of schools and attributes (SAT scores, majors offered, region, cost, public/private/for profit, etc.)
- “Where it Pays to Attend College” ([Kaggle.com](https://www.kaggle.com)) obtained from ([\*Wall Street Journal\*](#)), based on Payscale, Inc. ([College Salary Report Methodology](#)): Article reporting salaries of graduates, salaries by major, etc.

## Merge

- Significant cleaning, reformatting to match sets on college name
- String manipulation, removal of hyphens, abbreviations, region names, etc.

## More

- Sklearn’s provided “Iris” data

# Models

Scikit-learn classifiers that predict probabilities (predict\_proba implemented)

Multiclass logistic regression ([sklearn.linear\\_model.LogisticRegression](#))

- 85% accuracy on College data

Random Forest ([sklearn.ensemble.randomforestclassifier](#))

- 65% accuracy on College data

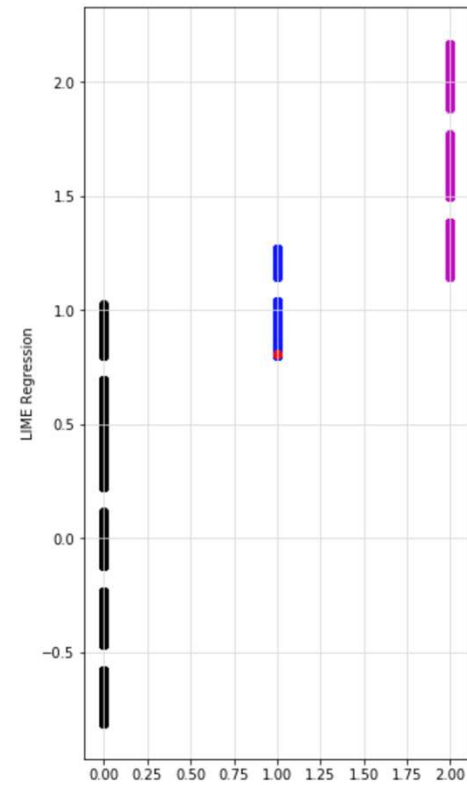
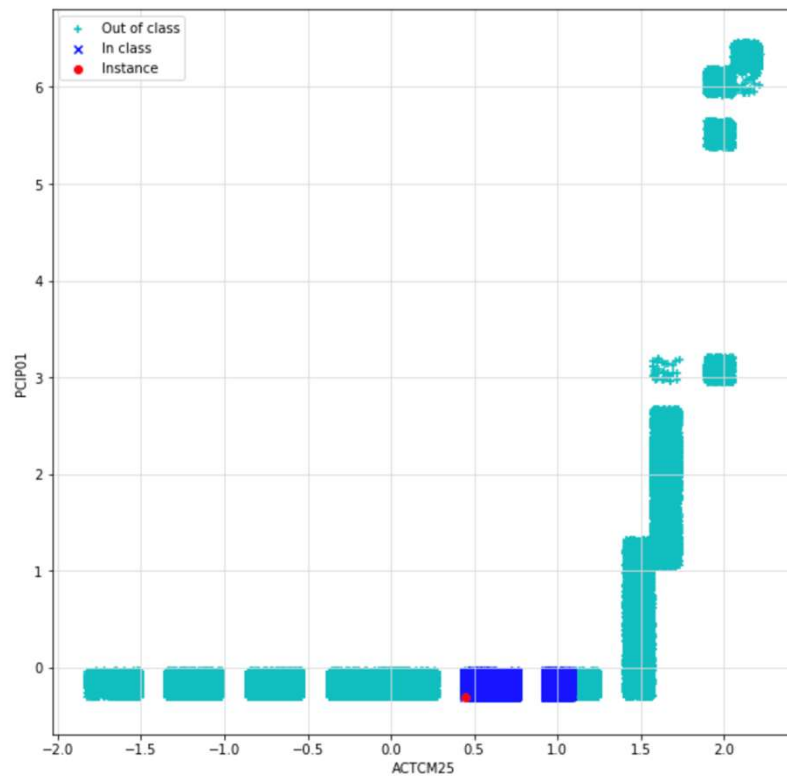
Decision tree ([sklearn.tree](#))

- 65% accuracy on College data

Models not tuned for improved accuracy (default settings)

# Demo

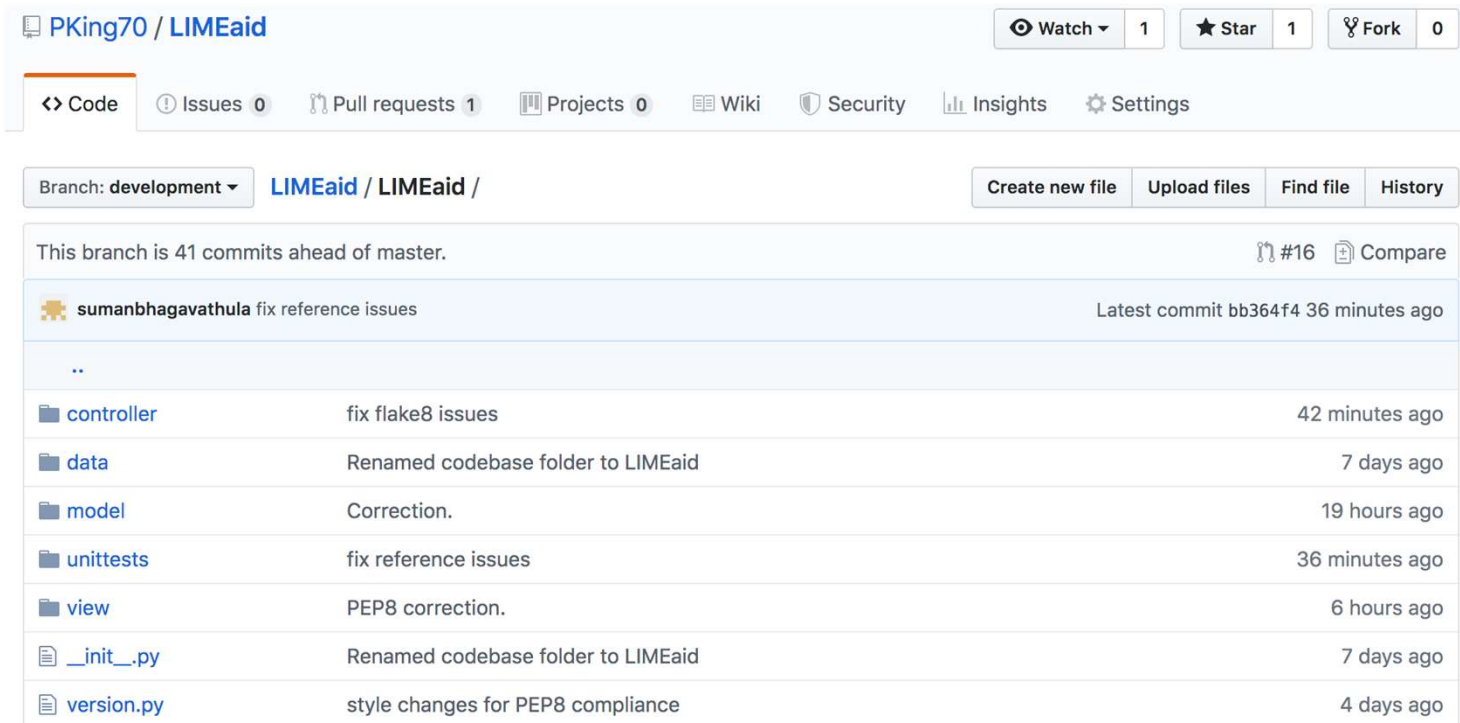
## Education and Iris data





# Project Structure

- [LIMEaid Github Repo](#) based on [Shablona](#)
- Used Model-View-Controller (MVC) codebase architecture



The screenshot shows the GitHub repository for PKing70 / LIMEaid. The repository has 1 watch, 1 star, and 0 forks. The main navigation bar includes links for Code, Issues (0), Pull requests (1), Projects (0), Wiki, Security, Insights, and Settings. The current branch is development, which is 41 commits ahead of master. A table of recent commits is displayed, showing the author, commit message, and time since the commit.

Commit Message	Time Ago
sumanbhagavathula fix reference issues	Latest commit bb364f4 36 minutes ago
..	
controller fix flake8 issues	42 minutes ago
data Renamed codebase folder to LIMEaid	7 days ago
model Correction.	19 hours ago
unittests fix reference issues	36 minutes ago
view PEP8 correction.	6 hours ago
__init__.py Renamed codebase folder to LIMEaid	7 days ago
version.py style changes for PEP8 compliance	4 days ago

# Future Work

API support for data acquisition to support dynamic features:

- College Scorecard ([data.gov](https://data.gov)) currently published with new features and data dictionary yearly

Model tuning for examples:

- Currently using defaults, could improve accuracy > 65%

Modify penalty for number of coefficients

Improve method to generate random samples

More data type and model support: image data, NLP support, support for model objects beyond sklearn classifiers

Questions?