Chronic Absenteeism Rate Prediction (CARP)

Coursera Advanced Data Science with IBM Specialization Capstone Project

Iver Band
December, 2019

Agenda

- Use Case
- Source Data Sets
- Architectural Overview
- Data Quality Assessment
- Data Visualization
- Data Exploration
- Initial Feature Engineering
- Model Performance Indicator
- Core Concepts
- Algorithms
- Frameworks
- Feature Engineering Experiment
- Model Performance Evaluation
- Conclusion

Chronic Absenteeism Rate Prediction (CARP) Use Case

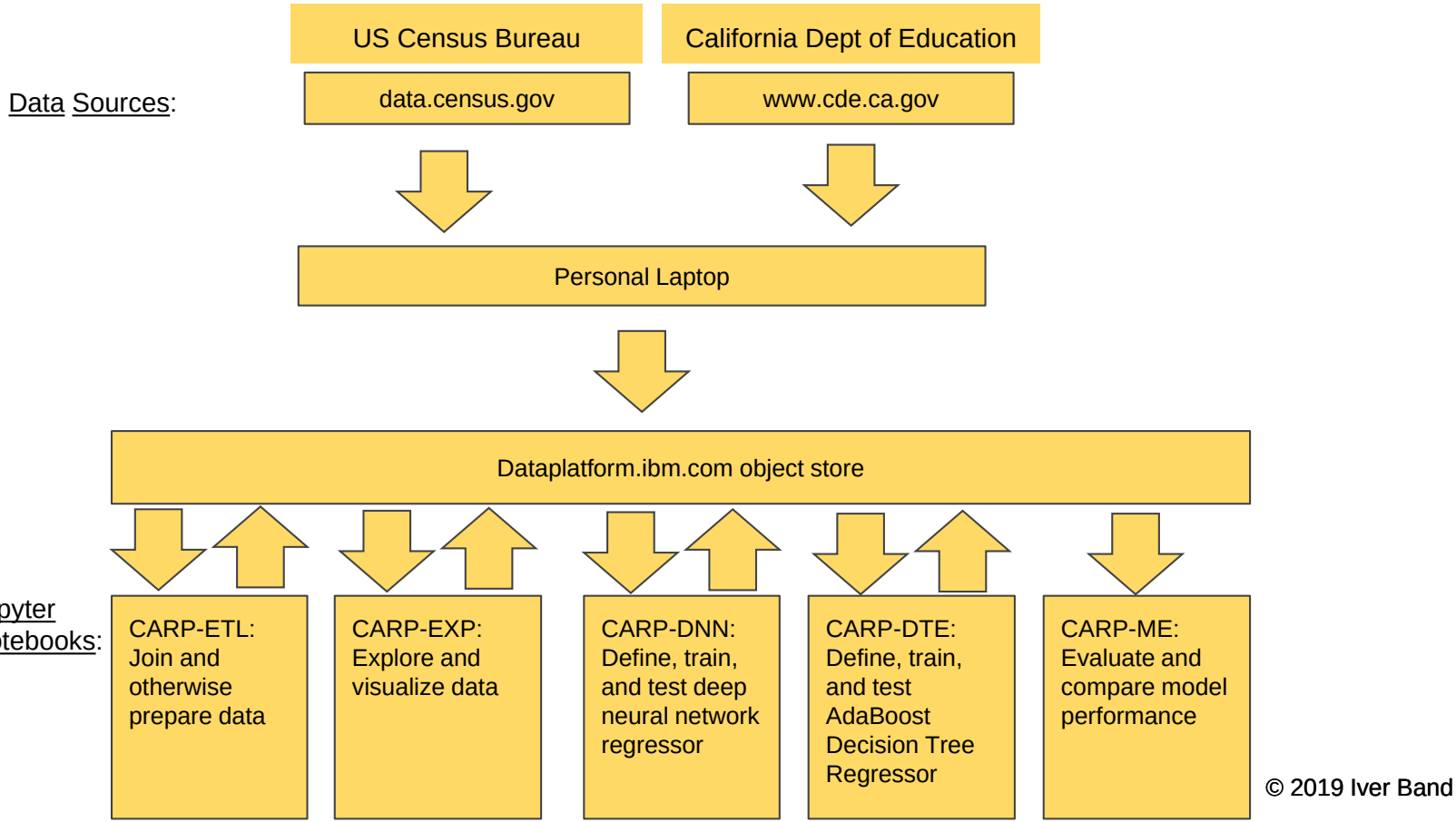
- Chronic absenteeism occurs when a student in grades K-12 misses 10% or more of the school year for any reason
- It is a strong predictor of low academic achievement.
- CARP is for data scientists supporting educational and social services administrators and policymakers in answering the following questions:
  - What demographic factors predict the rate of chronic absenteeism?
  - Can we use demographic data to predict chronic absenteeism rates?

Source Data Sets

- 2018 Chronic Absenteeism Data from California Department of Education
  - Counts of total students and chronically absent K-12 students by census tract in Los Angeles County, California, USA
- US Census Bureau American Community Survey 2013-2017
  - Statistics on potential predictors of 2018 chronic absenteeism rates:

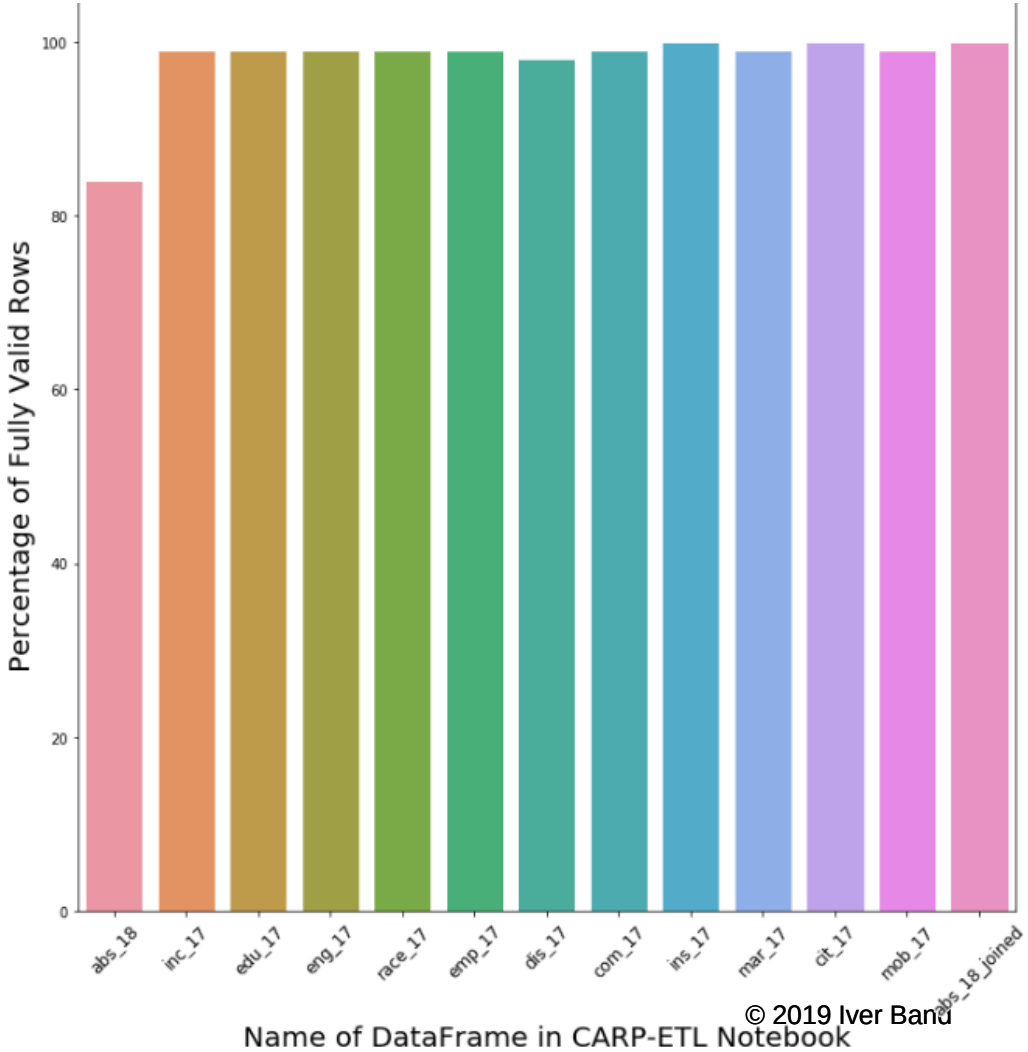
Median Income	Employment Status
Race	Length of Commute
Educational Attainment	Disability
Geographic Mobility	Marital Status
Health Insurance Status	Citizenship and Nativity
English Language Mastery	

Architectural Overview



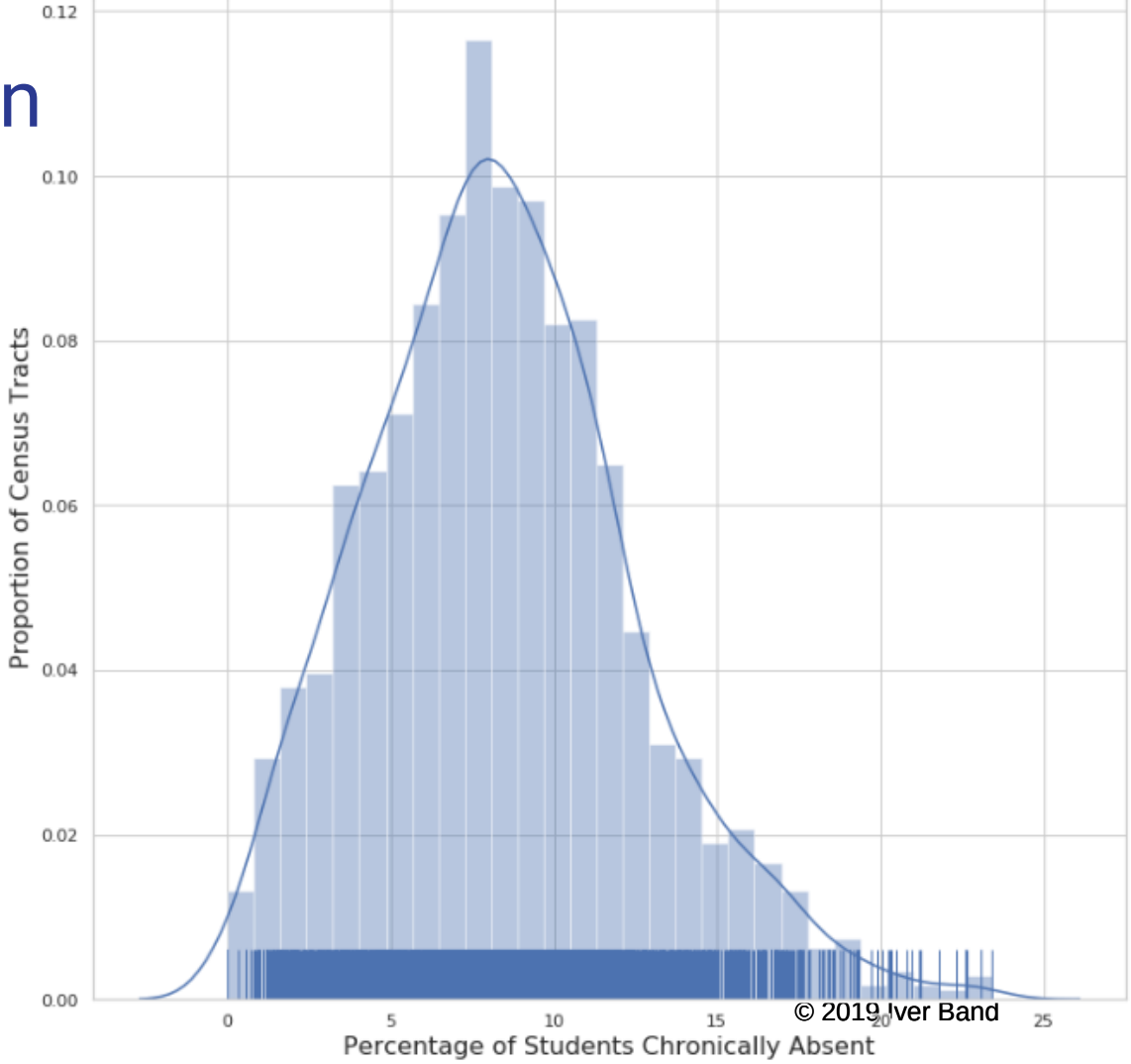
Data Quality Assessment

- Determined percentage of rows with missing or non-numeric data in
  - All 11 input data sets prior to cleansing
  - The joined data set
- Primary issue is 84% coverage of LA County census tracts,
- Data about 2158 census tracts was available for modeling



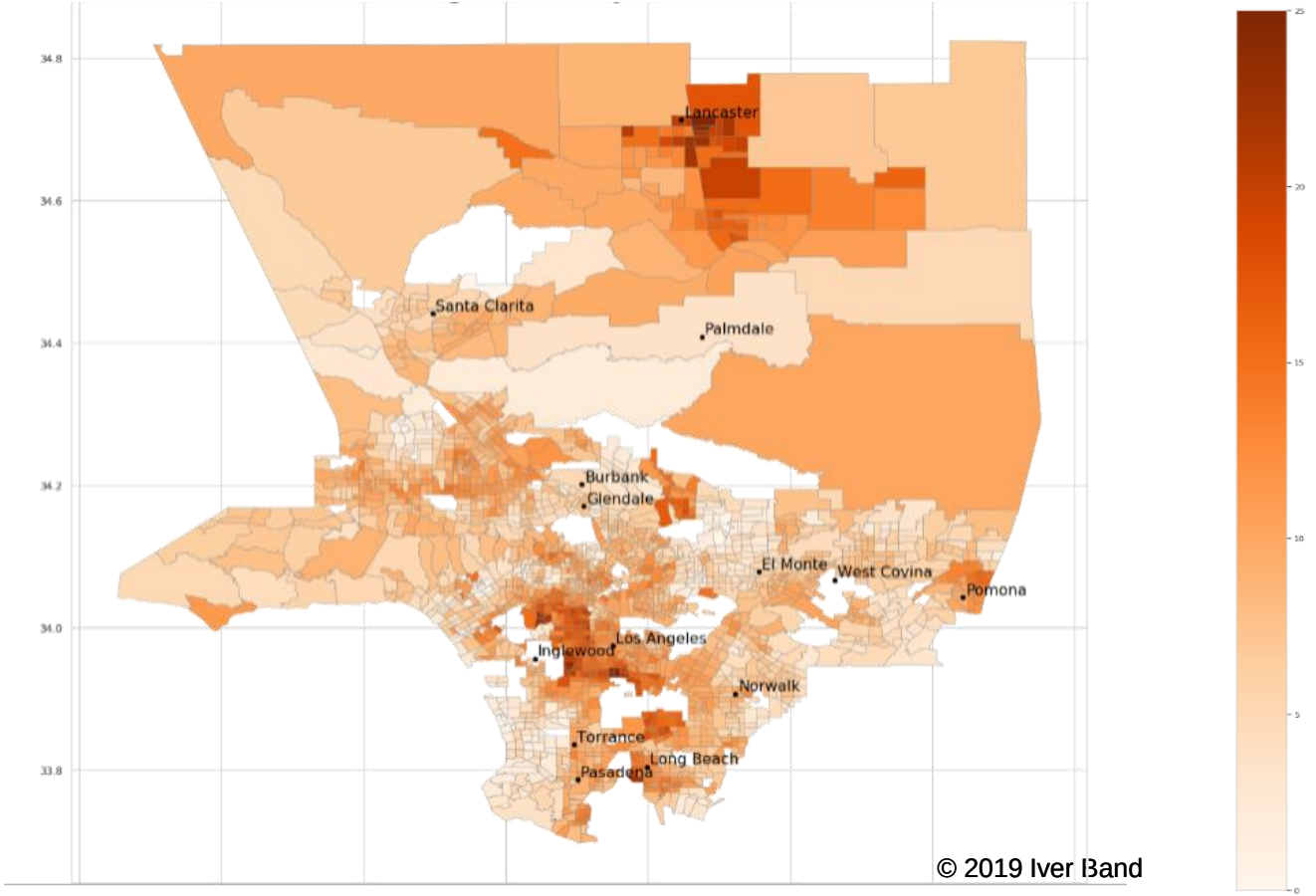
Data Visualization

- Visualized distribution of chronic absenteeism percentages with
  - Histogram
  - Kernel Density Estimate
  - Rug plot
- Distribution is roughly normal, with positive skewness



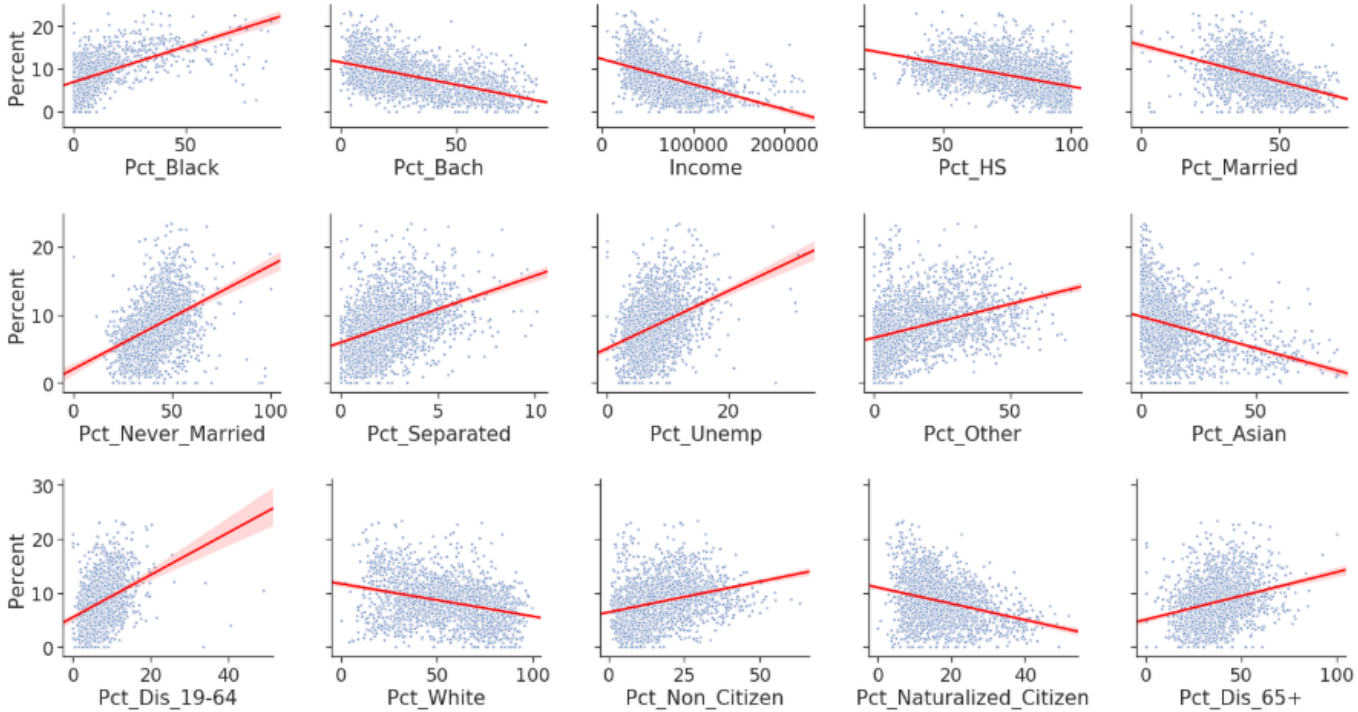
Data Visualization

- Visualized geographic distribution of chronic absenteeism rates with a choropleth map
- Adjacent or close census tracts tend to have similar rates



Data Exploration

- Visualized relationship of top 15 correlates per r^2 score using pair plots with regression lines
- Read left to right, then top to bottom
- Race, educational attainment, marital status and income are the strongest predictors
- Prediction strength declines sharply after the top five predictors



Initial Feature Engineering

- Converted counts into percentages in
  - California Department of Education chronic absenteeism data set
  - Nearly all of the American Community Survey data sets
- Joined all twelve data sets by six-digit US census tract number, which is unique with each county
- For deep neural network model only
  - Calculated r^2 for each predictor/target variable pair
  - Input only the top fifteen predictors
  - Scaled all data to have a mean of zero and a standard deviation of 1

Model Performance Indicator

Coefficient of Determination (r^2) Formula

- Measures the extent to which the predicted rates reflect the total variability of the actual rates
- Disproportionately penalizes large errors
- Prevents errors with opposing signs from cancelling each other out

$$1 - \frac{SS_{RES}}{SS_{TOT}} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Core Concept: Neural Network

- Feed forward: prediction by computing the value of each node, layer-by-layer, based on multiplying the nodes in the previous layer by a set of *weights*, adding the products and a *bias,* and applying a nonlinear *activation function*
- *Back propagation:* adjusting the weights and bias by differentiating the feed forward function to determine its slope, and applying *gradient descent* (next slide).

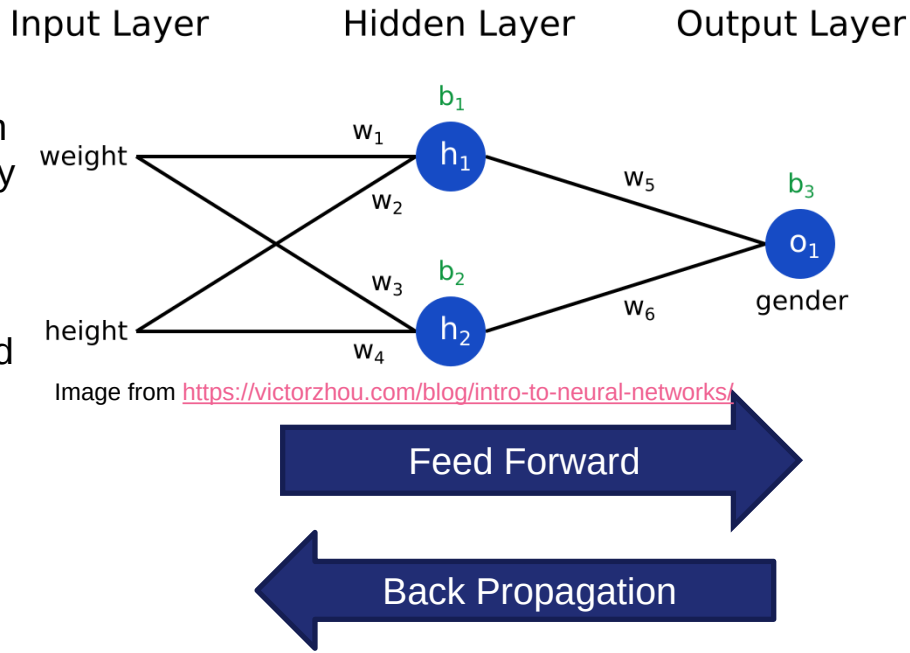
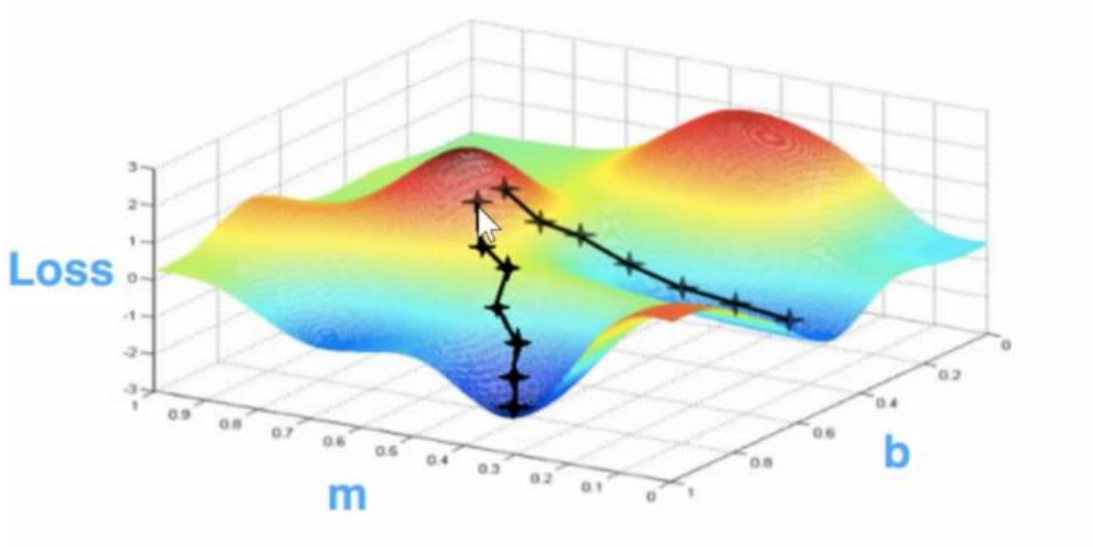


Image from https://victorzhou.com/blog/intro-to-neural-networks/

Core Concept: Gradient Descent for Neural Network Model Training

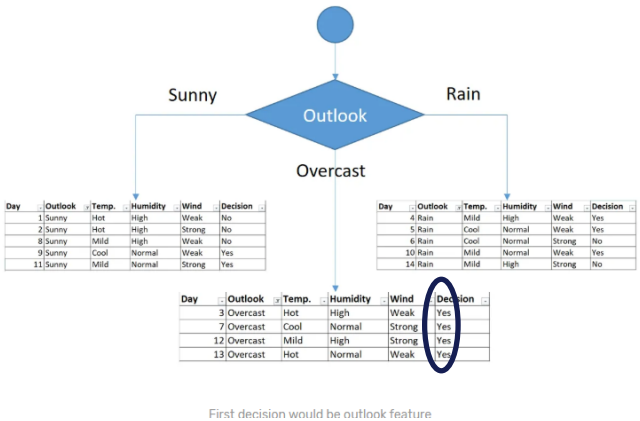
- Imagine trying to find the deepest valley in a deep fog
- Altitude is *loss function,* which compares the predicted value to the actual value
- The size of the step you take is the *learning rate*
- Descending the valley is *gradient descent*
- Your location in n-dimensional space is determined by using model weights and the loss function as coordinates
- The model-fitting algorithm repeatedly determines the gradient (slope) of the loss and adjusts model weights and biases to descend the gradient

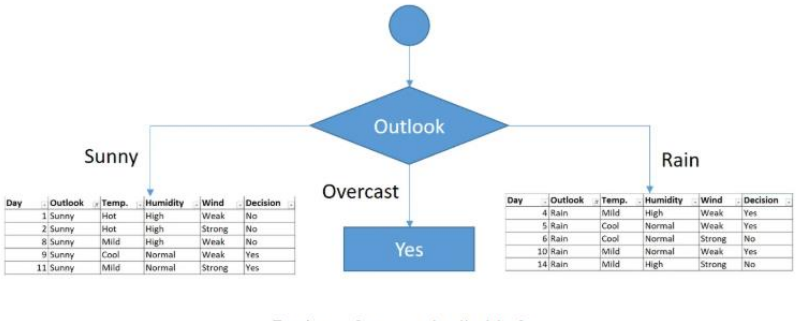


Text and picture adapted from
https://mc.ai/stochastic-gradient-descent-in-plain-english/

Core Concept: CART* for Decision Tree Modeling

Do I play golf today? Yes or No?

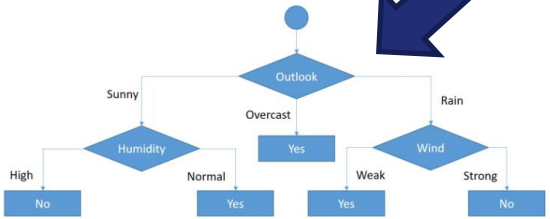




...

Build next leaf with decision criteria with lowest Gini score:

$$Gini = 1 - \Sigma (Pi)^2 \text{ for } i=1 \text{ to number of classes}$$



*Classification and Regression Trees

Adapted from: https://sefiks.com/2018/08/27/a-step-by-step-cart-decision-tree-example/

Algorithms

Feed-Forward Neural Network

- Dimensions of each layer
  - 15-->11-->7-->1
- RELU Activation for all nodes
  - f(x) = 0 if x <=0
  - f(x) = x otherwise
- Adam Optimizer
  - 0.01 learning rate
  - 0.001 decay
  - Mean squared error loss metric
- Batch size = 12
- 25-40 epochs (passes through data)
- Early stopping for training runs that don't converge
- Attributes chosen after wide experimentation

Decision Tree Ensemble

- AdaBoost with decision tree as weak learner
  - Repeatedly builds decision trees, weighting them to improve weakest predictions
  - Squared error loss metric
- Method chosen after experimentation with linear regression and standalone decision tree regression, as well as other decision tree ensembles, such as random forest

Frameworks

All Python-based

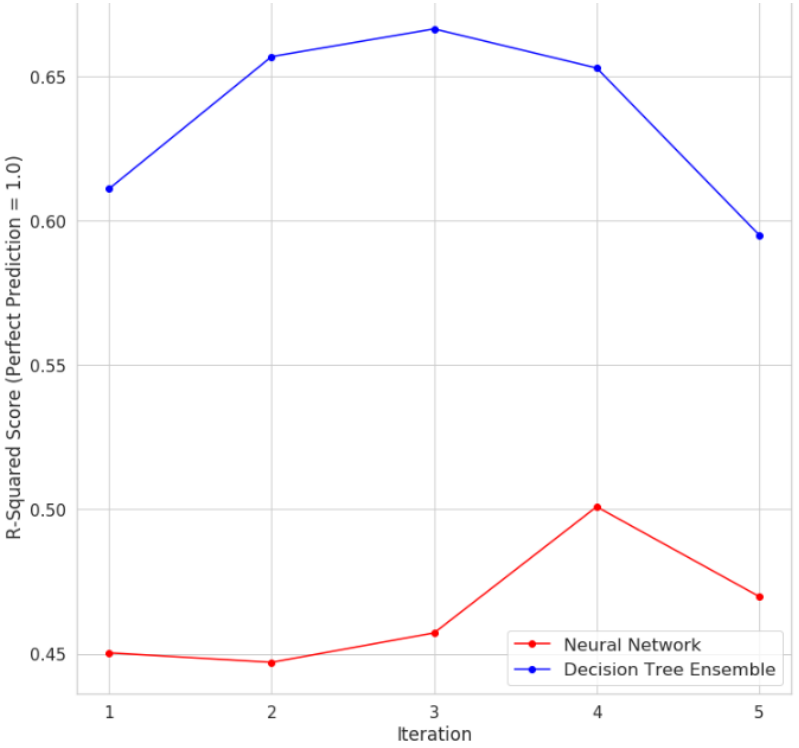
Purpose	Frameworks
Interactive Development	Jupyter within IBM Watson Studio
Data Manipulation and Computation	Numpy, Pandas, Geopandas
Data Visualization	Matplotlib, Seaborn, Descartes
Neural Network Modeling	Keras with TensorFlow backend
Data Scaling, Model Scoring and Decision Tree Ensemble Modeling	Scikit-Learn

Feature Engineering Experiment

- Choropleth map shows adjacent census tracts tend to have similar chronic absenteeism rates
- Many adjacent or nearby census tracts in LA County appear to have similar or sequential census tract numbers
- Therefore, would including the six-digit census tract number as a predictor variable, in addition to its role as a key, improve model performance?
- Unfortunately not. The census tract number is among the weakest of the predictor variables examined, based on
    - The pairwise r^2 score
    - No performance change for the Decision Tree Ensemble model

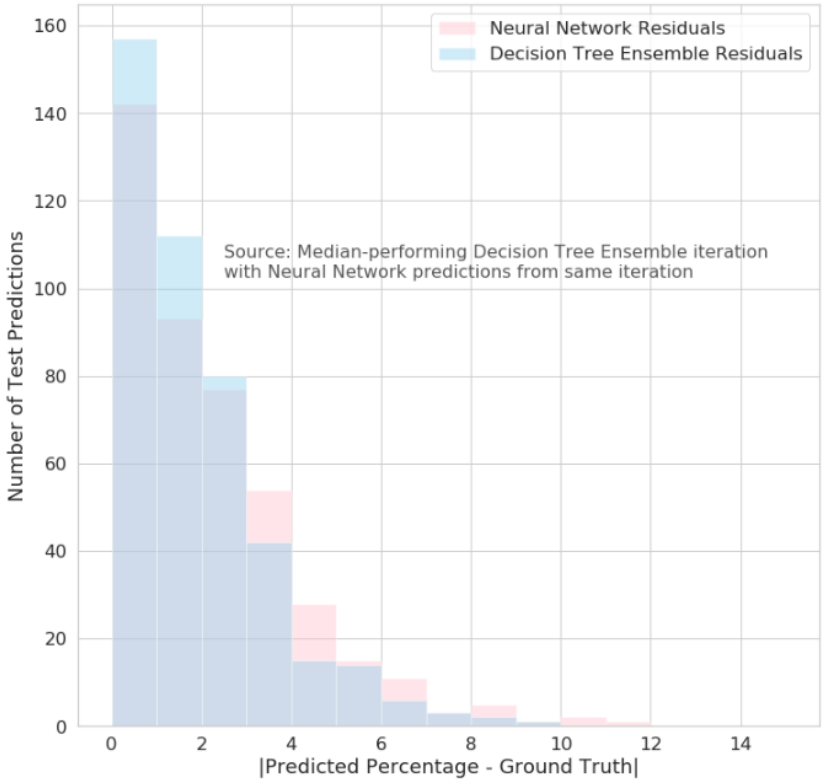
Model Performance Evaluation

- Tested both models with five rounds of randomized shuffle-split with 80/20 train-test ratio
- Chose randomized shuffle-split over cross-validation to retain finer control over size of split
  - Before tuning, got good performance only with 90/10 train-test split
  - Future work could use cross-validation
- **The Decision Tree Ensemble is clearly a stronger predictor than the Neural Network**
- To avoid outliers, chose iteration with median performance of stronger model for further evaluation



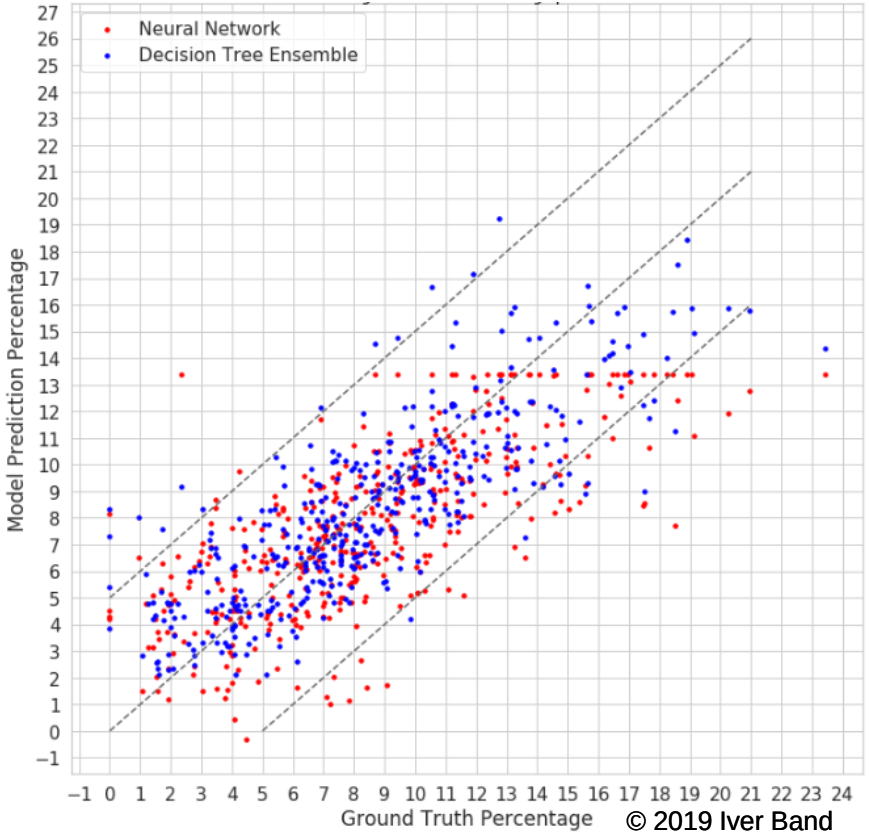
Model Performance Evaluation

- Compared residuals of predictions using superimposed histograms
- For about 90% of test cases in its median-performing iteration, Decision Tree Ensemble predicted chronic absenteeism rates within 5%



Model Performance Evaluation

- Used a scatter plot with an identity and +/- 5% lines to visualize deviations of predicted chronic absenteeism rates from actual values
- Result is consistent with residual histograms



Conclusion

- What demographic factors predict the rate of chronic absenteeism?
  - ***For the years studied, race, educational attainment, marital status and income are the strongest predictors of chronic absenteeism in LA County, California, USA***
- Can we use public demographic data to predict chronic absenteeism rates?
  - ***Yes, for the years studied, well enough to identify areas with highest future risk and, perhaps, implement mitigating measures***
- Opportunities for future work
  - Expand scope to include additional geographic areas, years, and predictor variables
  - Further automate data extraction and transformation
  - Consider additional algorithms, e.g. PCA, XGBoost, Polynomial Regression
  - Explore outliers where rate is more or less than expected
  - Explore effects of programs to mitigate chronic absenteeism
- Links

  https://github.com/IverBand/carp

  https://www.coursera.org/specializations/advanced-data-science-ibm