



# Interpretable Deep Credit Risk Ranking

Kyle Grove

Chief Data Scientist, US Financial Industry

Teradata Corporation

# Interpretable Deep Credit Risk Ranking

Disclaimer: the efforts described in this presentation are solely Teradata's laboratory efforts and no partnership with other business entity is implied.

## Contents

- Domain of Credit Risk Analytics
- Deep and Broad Sequence Learning
- Deep and Broad Default Prediction: A Case Study
- Model Interpretability

# Domain of Credit Risk Modeling

## Credit Risk

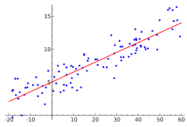
- “It seems to be a law of nature, inflexible and inexorable, that those who will not risk cannot win.” John Paul Jones
- Lenders seek to estimate the credit risk presented by prospective borrowers, in order to:
  - **Decide whether to extend credit.**
  - **Decide at what rate to extend credit.**

## Desiderata of Credit Risk Models

- **calibration** — the model needs to consistently rank high-risk customers higher than low-risk customers.
- **accuracy** — the model needs to guide profitable lending decisions; the value of a model parameter reflects the parameter value in the true model.
- **interpretability** — model parameters that drive predictions need to be interpretable by regulators and stakeholders.

# Statistics vs. Machine Learning

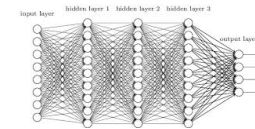
- Statistics yields a set of interpretative techniques for understanding the behavior of a complex system.
- Machine Learning yields a set of predictive techniques for building machines that make accurate predictions on data.



regression



decision trees and  
their kin



deep neural nets

## Accuracy vs Interpretability

- **accuracy** — the model needs to guide profitable lending decisions; the value of a model parameter reflects the parameter value in the true model.
- **interpretability** — model parameters that drive predictions

**Feature combination, nonlinearity, and highly granular features** effects may *boost accuracy*...

...but **feature combination, nonlinearity, and feature** will *challenge interpretability*.

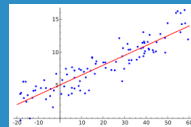


# The Regulated Modeler's Dilemma

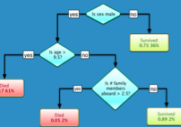
Everything the light touches is explainable.



regression



Markov machines

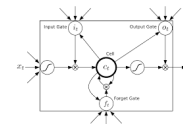
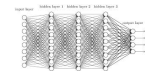


decision trees

But what about that shadowy place?

That's beyond our borders. You must never go there, Simba.

deep learning

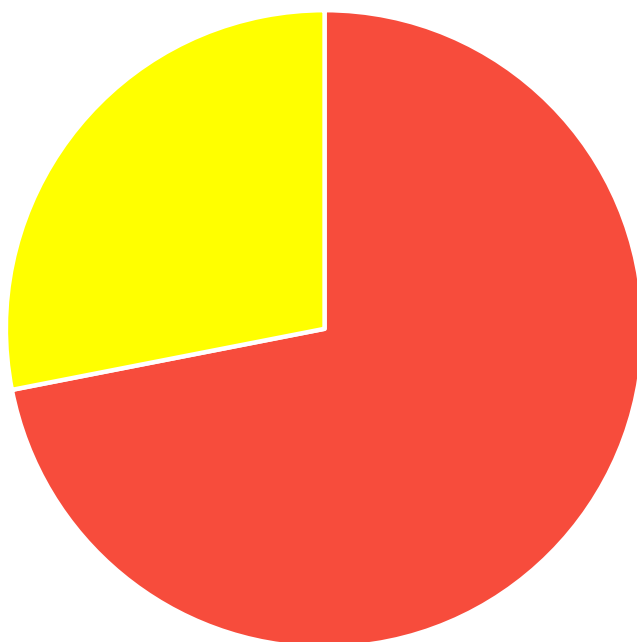


decision tree forests



**The regulated modeler's dilemma:  
any gain in accuracy through  
cleverness is only useful if the  
cleverness is obvious.**

## So, Are Banks Generally Able at Present to Use Advanced ML in Lending Decisions?



■ no ■ no, but in yellow

## So, Are Banks Safe with their Incumbent Credit Models?

There is a strong relationship between income and having a scored credit record.

**26** million Americans credit invisible +  
**19** million Americans credit unscorable

[consumerfinance.gov/f/201505\\_cfpb\\_data-point-credit-invisibles.pdf](https://consumerfinance.gov/f/201505_cfpb_data-point-credit-invisibles.pdf)

## So, Are Banks Safe with their Incumbent Credit Models?

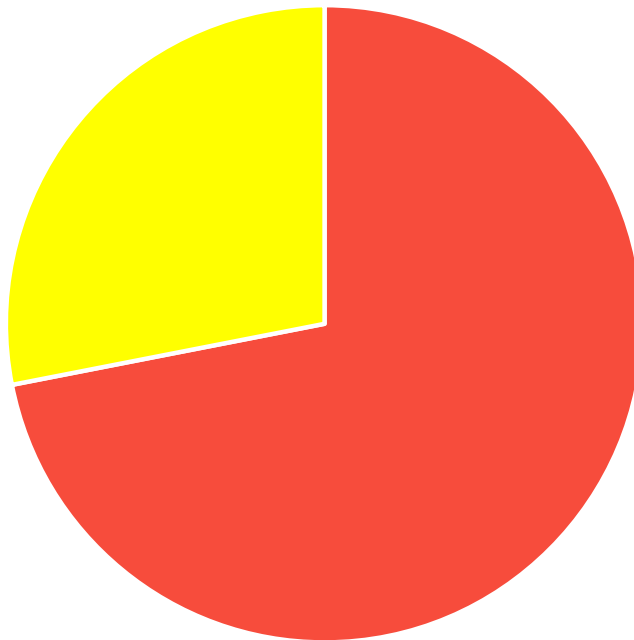
**11482**      **Federal Register** / Vol. 78, No. 32 / Friday, February 15, 2013 / Rules and Regulations

---

(a) *Discriminatory effect.* A practice has a discriminatory effect where it actually or predictably results in a disparate impact on a group of persons or creates, increases, reinforces, or perpetuates segregated housing patterns because of race, color, religion, sex, handicap, familial status, or national origin.

<https://www.hud.gov/sites/documents/DISCRIMINATORYEFFECTRULE.PDF>

Does the regulatory framework amnesty lenders who simply avoid protected-class demographic data in their models?



■ no ■ no, but in yellow

# So, Are Banks Safe with their Incumbent Credit Models?

**11482**      **Federal Register** / Vol. 78, No. 32 / Friday, February 15, 2013 / Rules and Regulations

---

(a) *Discriminatory effect.* A practice has a discriminatory effect where it actually or predictably results in a disparate impact on a group of persons or creates, increases, reinforces, or perpetuates segregated housing patterns because of race, color, religion, sex, handicap, familial status, or national origin.

## So, Are Banks Safe with their Incumbent Credit Models?

The net effect is pressure on financial institutions to enhance not just the accuracy of credit models, but also the fairness of credit models.

- Proportionality – model predictions avoid disparate impact
- Transparency – model predictions are interpretable

## Proportionality Techniques (Future Work)

- Models to detect and intercept disparate impact before production
- Structured regularization to minimize disparate impact

## Proportionality Techniques (Here & Now)

- Tools and techniques for model interpretability



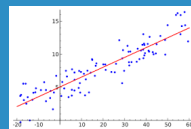
# The Regulated Modeler's Dilemma



0.5 AUC

coin  
flipping

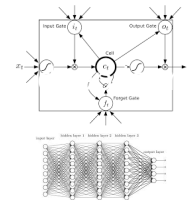
regression



Markov  
machines



decision  
trees



deep learning

decision tree forests



Accuracy

How do we cut the Gordian knot of  
uninterpretable accuracy?

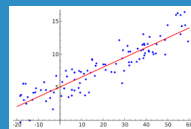
# The Regulated Modeler's Dilemma



0.5 AUC

coin  
flipping

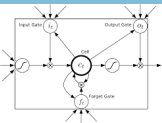
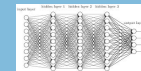
regression



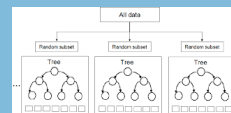
Markov  
machines



decision  
trees



deep learning



decision  
tree forests

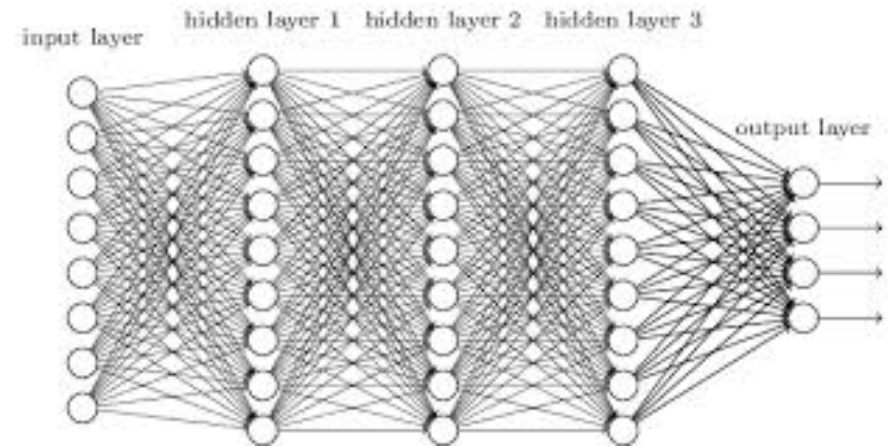
Accuracy

**Strategy: operationalize more  
sophisticated model classes and  
develop interpretation methods and  
tools**

# Deep and Broad Sequence Modeling

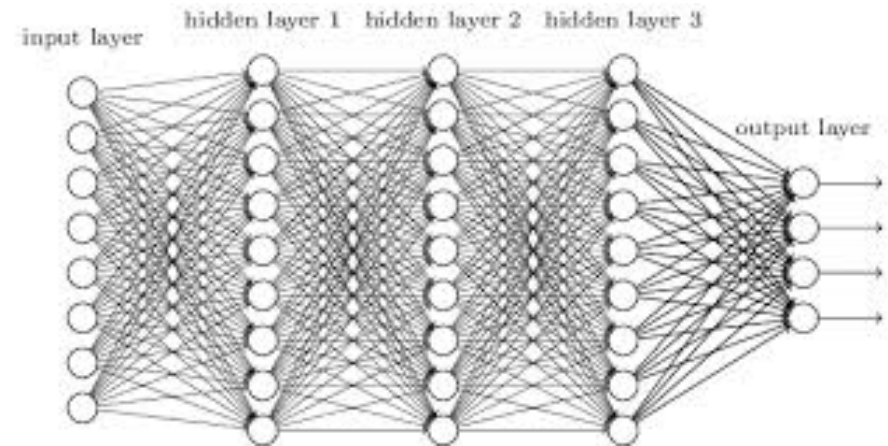
## Deep Neural Nets

- Neural Networks are a powerful model class that utilizes abstract computing units referred to as *neurons* or *nodes*.
- A neuron outputs a numerical *activation* that is a weighted function of its inputs.
- A multi-layer neural net learns these weights by using *backpropagation* aka the chain rule to partition error backwards through the network.



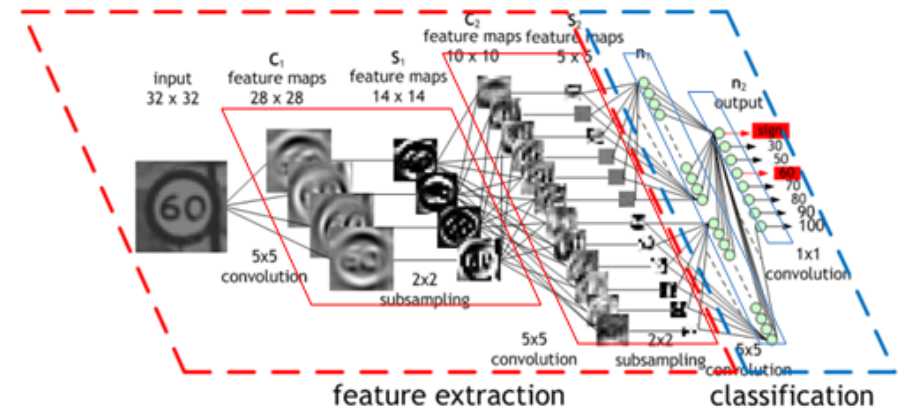
## Deep Neural Nets

- Neural Networks were invented in the 1950s but only recently became useful with some algorithmic innovations + massive amounts of compute and training data.
- By stacking many layers of neurons into a neural network, we hope to build a *universal function approximator* that can learn complex high-dimensional patterns..



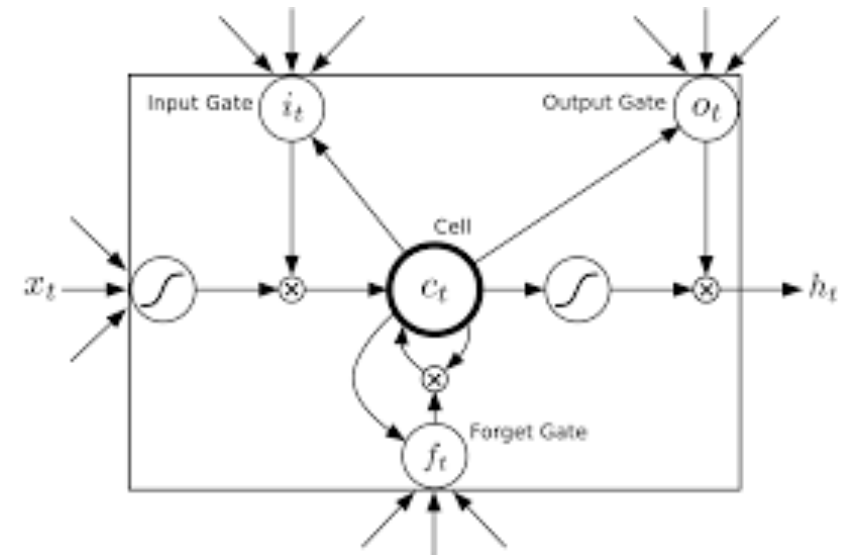
# Convolutional Neural Nets

- **Idea!** Employ standard neural networks to learn from 2D images.
- **Problem!** Learning separate parameters for each pixel does not scale.
- **Solution!** Use *convolution* over multiple *receptive* fields to learn specific image features that generalize across the entire visual field.



## Long Short Term Memory

- **Idea!** Employ standard neural networks to learn from panel data.
- **Problem!** Learning separate parameters for each time tick does not scale.
- **Solution!** Use *recurrence* and *gating* to learn features that generalize across the time course.



## Gradient Boosted Forest

- **Idea!** Combine the best part of decision tree learning (discontinuities) with the best part of regression modeling (additivity).
- **Problem!** How to do it.
- **Solution!** In sequence, fit each decision tree to the *residuals* from all previous stages.



# Deep and Broad Default Prediction

# Freddie Mac Single Family Loan Data

## Single Family Loan-Level Dataset

Data on fully amortizing, 30-year fixed-rate mortgages originated between 1/1/99 and 3/1/14

### Non-panel origination data

- 20M+ records
- **predictors:** includes factors which banks consider in extending a home loan, including *credit score*, *first time homebuyer*, *PMI%*, *DTI ratio*, and *loan-to-value %*.

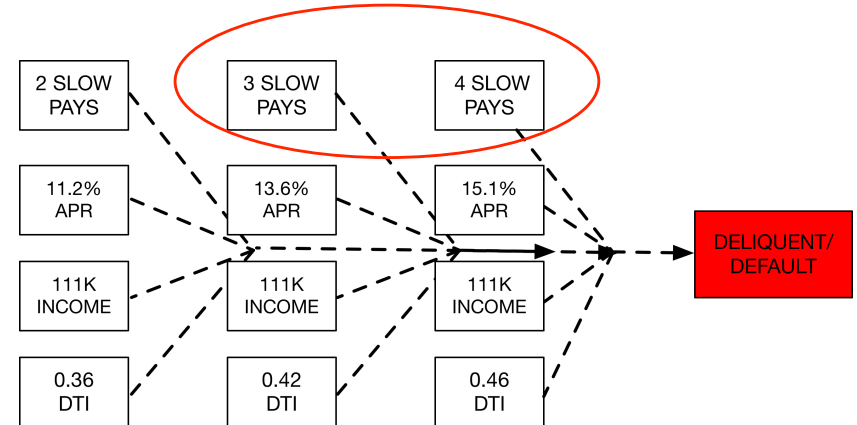
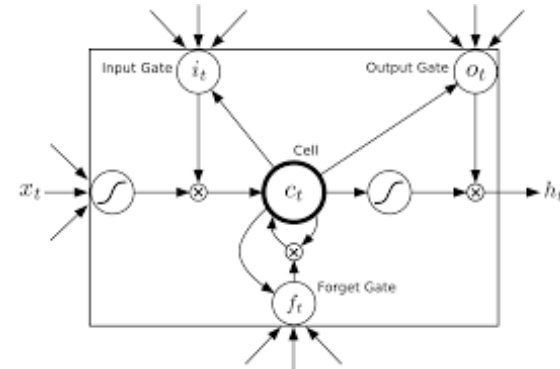
### Panel performance data

- 1B+ records
- **outcomes:** monthly loan performance data through 9/30/14



# Deep and Broad Credit Risk Prediction

- A Long Short Term Memory model learns to make predictions of future events based on knowledge of the most recent event plus patterns from previous context.
- Gradient Boosting learns constellations of thresholds that predict well on novel data.

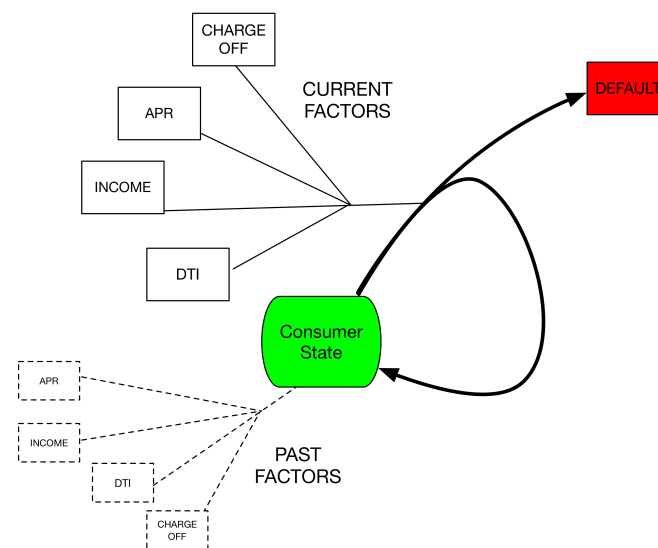


# Applying Deep and Broad Models to Predict Delinquency in Freddie Mac Single Home Dataset

Utilized an ensemble classification model (single layer LSTM + XGBoost) that balances between present factors and past factors.

Fit the ensemble models to the training partition to predict delinquency.

Forecast the ensemble into the out-of-time validation months, and measured AUC of the ensemble against the validation data.



## Applying Deep and Broad Models to Predict Delinquency in Freddie Mac Single Home Dataset

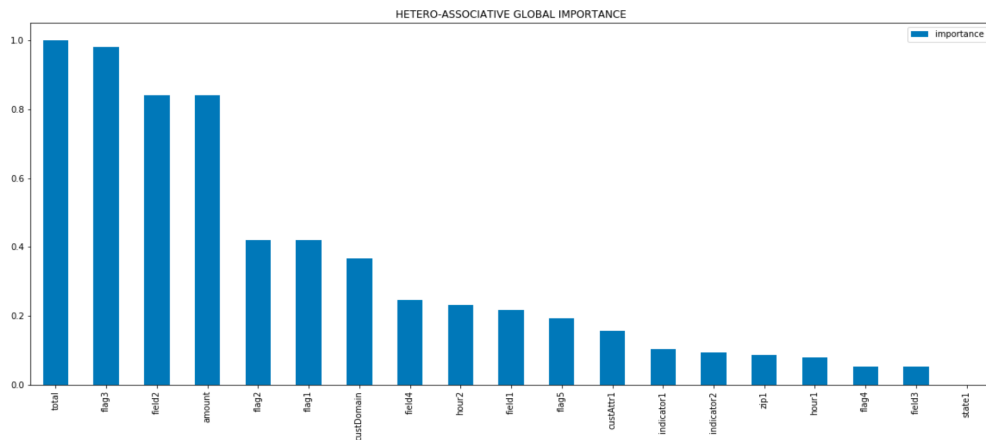
| model                            | AUC  |
|----------------------------------|------|
| logistic regression              | 0.72 |
| LSTM, CNN, &<br>XGBoost ensemble | 0.89 |

# Model Interpretability

# Model Interpretability Desiderata

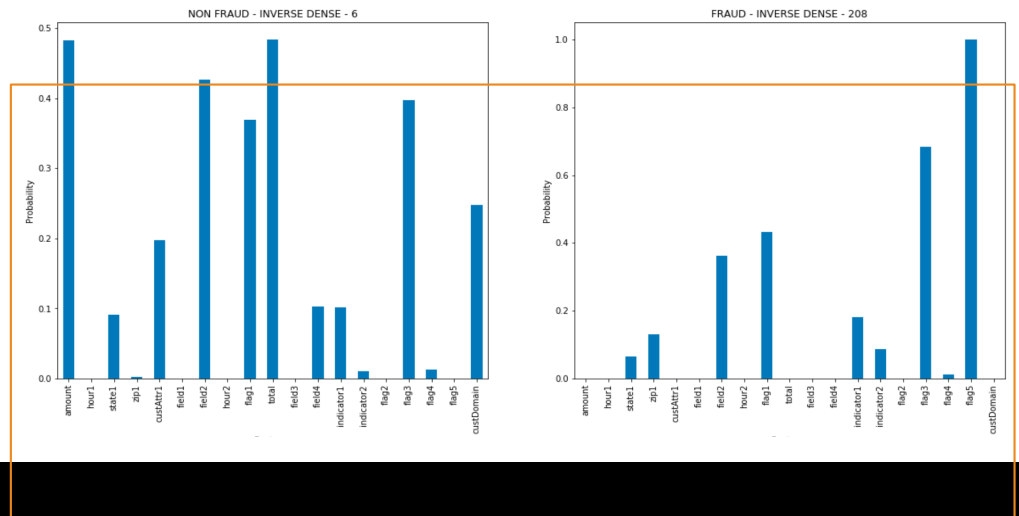
## Global Interpretability

Convey the impact of parameters across predictions in the aggregate



## Local Interpretability

Convey the impact of parameters on individual predictions



## Model Interpretability Strategies

- Augmented Data Proxy Models – the uninterpretable model to be operationalized is interpreted by proxy models fit to a local, synthetically generated neighborhood of the prediction
- Directly Fit Proxy Models – the uninterpretable model to be operationalized is interpreted by proxy models fit to a subset of training data relevant to the prediction
- Model Scoring – the uninterpretable model to be operationalized is scored by a procedure so as to be interpreted via the scores yielded by the procedure
- Model Distillation – the uninterpretable model transmits compressed knowledge to an interpretable model to be operationalized



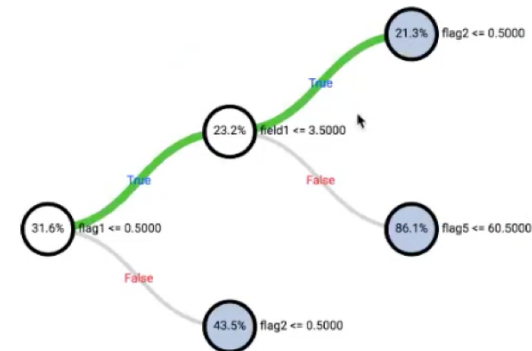
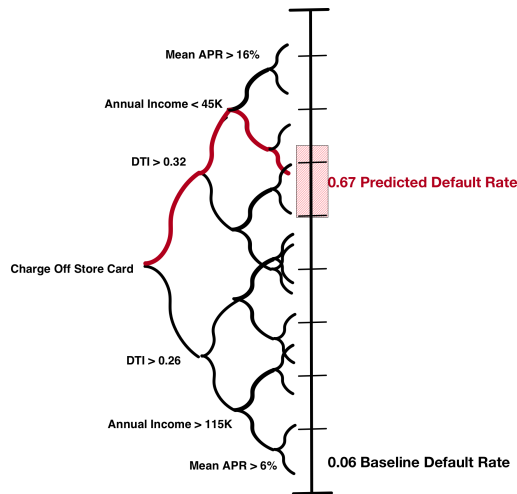
## Augmented Data Proxy Models

- Example: LIME
- Strategy: Generate a set of synthetic data points from training data in the neighborhood of the prediction, and fit a local interpretable model to the prediction
- Pros: interpretable, black box, and intuitive
- Cons: may not natively scale to interactive speed, synthetic data points.

## Directly Fit Proxy Models

- Example: Modified LIME
- Strategy: Sample a set of relevant data points from training data in the neighborhood of the prediction, and fit a local interpretable model to the prediction
- Pros: interpretable, black box, fast and intuitive
- Cons: trades off time for space to leverage precomputation

# Decision Tree Journey Map



- Build a global constrained decision tree using modified LIME sampling as global model that can trace the feature path for a particular individual.
- Particularly useful for sequence data, can constrain earlier events to be at the root of the tree to drive the journey map intuition.


## Model Scoring

- Example: xgboost's feature\_importance, Shapley scoring
- Strategy: Utilize the results of a heuristic scoring procedure to stand in for the actual relevant parameter values.
- Pros: interpretable and intuitive. model free.
- Cons: the heuristic scores can lack fidelity to the parameters. Not necessarily black box.

# Shapley Scoring



| Western Conference |  | W  | L | Pct  | GB | Conf | Div  | Home | Away |
|--------------------|--|----|---|------|----|------|------|------|------|
| 1                  |  Warriors | 73 | 9 | .890 | -  | 46-6 | 15-1 | 39-2 | 34-7 |

| Western Conference |  | W  | L  | Pct  | GB | Conf  | Div  | Home | Away  |
|--------------------|--|----|----|------|----|-------|------|------|-------|
| 1                  |  Warriors | 67 | 15 | .817 | -  | 42-10 | 14-2 | 36-5 | 31-10 |

- Intuitively, what is the return when we add one player to the coalition?
- Utilizes the entire training set to determine feature share of superadditive interactions.

## Shapley Scoring

- Advantage 1: built from the actual training dataset
- Advantage 2: uses the entire dataset on the prediction (less prone to overfit)
- Disadvantage: Can be expensive to (1-time) compute for high dimensional data, but shortcuts exist.

## Model Distillation

- Example: Caruana-style Model Compression, Model Teaching, Co-Regularization
- Strategy: Use a complex model to teach a more interpretable model.
- Pros: Appealing in the regulated space where the interpretable model class may already be acceptable to regulators.
- Cons: Complex parameterizations may not translate between source and target model classes (e.g. a neural net learns interactions that may not be representable in a linear regression).

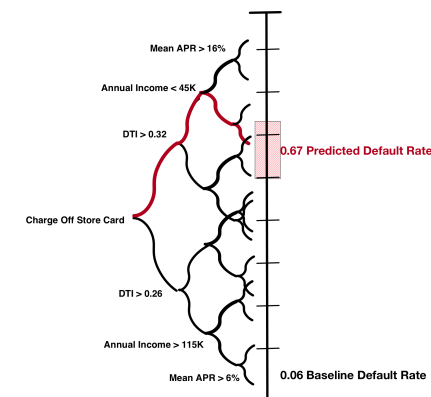
# Model Interpretability Tool

| caseid | prediction | charge off | ?late last month | DTI  | Annual Income |
|--------|------------|------------|------------------|------|---------------|
| 1123   | 0.67       | T          | T                | 0.36 | 114K          |
| 89765  | 0.18       | T          | F                | 0.22 | 55K           |
| 3827   | 0.02       | F          | F                | 0.13 | 65K           |
| 3426   | 0.01       | F          | F                | 0.15 | 35K           |
| 9238   | 0.03       | F          | F                | 0.16 | 88K           |
| 3654   | 0.07       | F          | F                | 0.20 | 127K          |

prediction case menu

| Factor                            | Risk |
|-----------------------------------|------|
| Charge Off Store Card             | +++  |
| ?Late Last Month Primary Card     | ++   |
| DTI > 34% Last Month              | ++   |
| >10 #Days Late Last Month Primary | ++   |
| DTI > 32% Today                   | ++   |
| Annual Income > 85K               | -    |
| Mean APR% < 8%                    | --   |

Predictive interpretability with variety of techniques



journey mapping with constrained decision tree technique



# Model Interpretability Tool

|        |       |        |       |           |        |        |       |       |             |
|--------|-------|--------|-------|-----------|--------|--------|-------|-------|-------------|
| amount | hour1 | state1 | zip1  | custAttr1 | field1 | field2 | hour2 | flag1 |             |
| 0.0    | 0.0   | 0.0    | 0.013 | 0.0       | 0.0    | 0.370  | 0.0   | 0.518 | ← Non-Event |
| 0.0    | 0.0   | 0.0    | 0.121 | 0.0       | 0.0    | 0.182  | 0.0   | 0.924 | ← Event     |

|       |        |        |            |            |       |       |       |       |             |
|-------|--------|--------|------------|------------|-------|-------|-------|-------|-------------|
| total | field3 | field4 | indicator1 | indicator2 | flag2 | flag3 | flag4 | flag5 |             |
| 0.0   | 0.0    | 0.0    | 0.119      | 0.017      | 0.0   | 0.363 | 0.016 | 0.0   | ← Non-Event |
| 0.0   | 0.0    | 0.0    | 0.201      | 0.060      | 0.0   | 0.231 | 0.029 | 1.0   | ← Event     |

