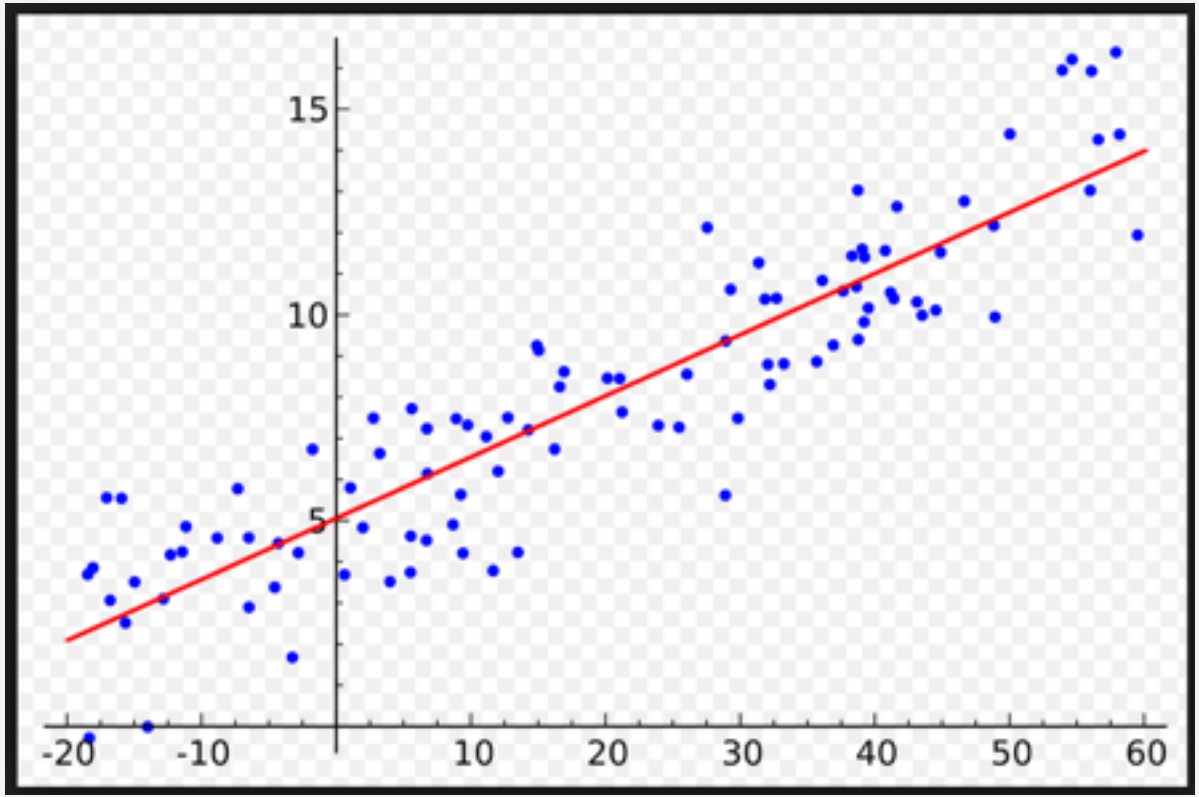
Interpreting Black-Box Models with Applications to Healthcare

Brian Lucena

In The Beginning…



Variables had Clear Interpretations

- Relationships were Linear
- Features were Interpretable
- Interactions were ignored (or manually added)
- Predictive Performance was OK

Then came Advanced ML

- Random Forests, Gradient Boosting, Deep NN
- Superior Predictive Power
- Capture High-Order Dependencies
- Lack the Simple Interpretations

Linear / Logistic Regression

- Unit increase always has same effect
  - Current Value of Feature is Irrelevant
  - Context of Other Features is Irrelevant
- Reality: These things matter!

Consider: How much does an additional 100SF add to a home price?

Three Reasons to Interpret Models

1. To Understand how Features contribute to Model Predictions - Build Confidence

2. To Explain Individual Predictions

3. To Evaluate the Consistency / Coherence of the Model

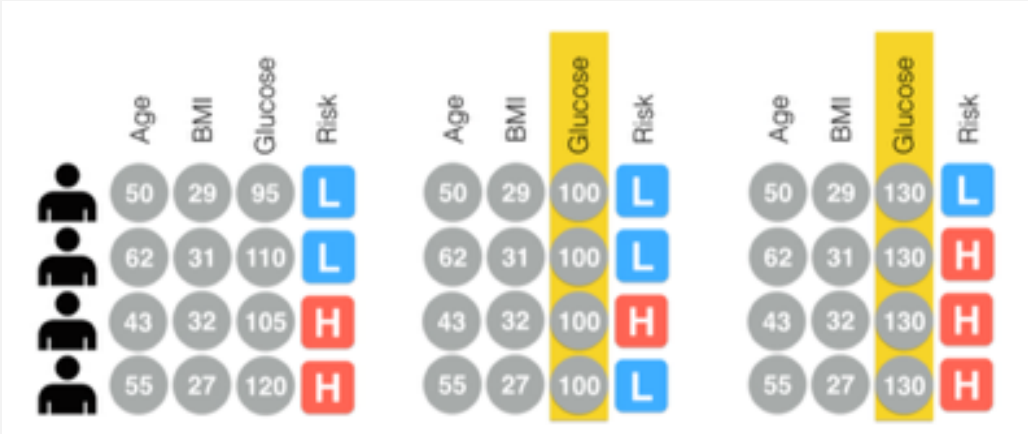
Some Work on Model Interpretation

- Partial Dependence (Friedman, 2001)
- Average the effect of a single variable, marginally and empirically.

Graphic: Krause, Perer, Ng. Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. ACM CHI **2016**

Some Work on Model Interpretation

- ICE-Plots (Goldstein et al, 2014)*
- Look at “trajectory” of each data point individually.



*Goldstein, Kapelner, Bleich, Pitkin. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. J*ournal of Computational and Graphical Statistics* (March 2014)

Some Work in Model Interpretation

- "Model-Agnostic Interpretability of Machine Learning" Ribeiro, Singh, Guestrin. https://arxiv.org/abs/1606.05386
- Create a "locally-interpretable" model in region of interest.
- Good reference document

ML-Insights Package (Lucena/Sampath)

- Idea: Pre-compute a “mutated” data point along each feature across a range of values
- Enables quick Feature Dependence Plots (ICE-plots) - model exploration / understanding
- Feature Effect Summary - more meaningful representation than “feature importances”
- Explain Feature Prediction - given two points, explain why model gave different predictions

ML-Insights Example 1

Example 1: Ames Housing Data*

- Predict housing prices from a subset of 9 variables.
- 2925 data points, split 70/30 train/test
- Fit two models: Gradient Boosting and Random Forest

*Dean De Cock, *Journal of Statistics Education* Volume 19, Number 3(2011), www.amstat.org/publications/jse/v19n3/decock.pdf

ML-Insights Example: Housing

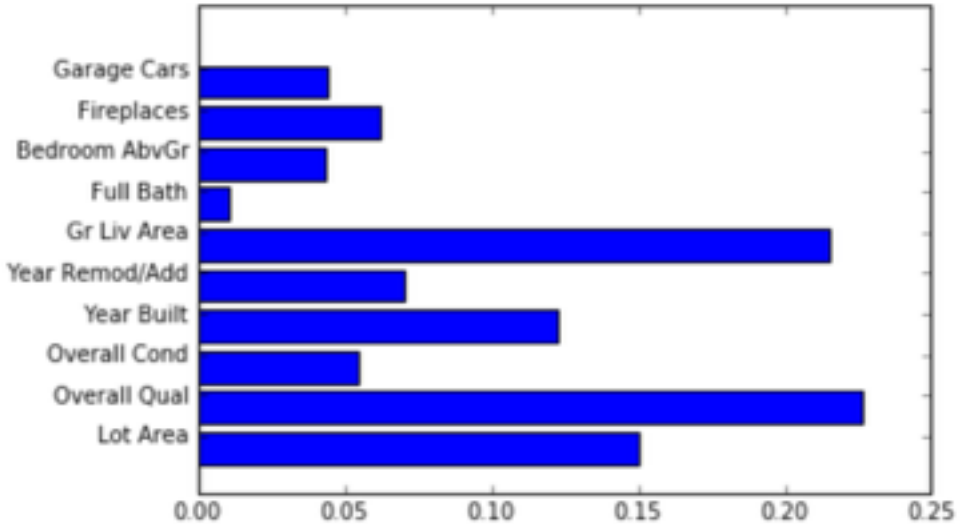
In [10]:
```
#Split the data 70-30 train/test

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
                                                    random_state=42)
```

In [11]:
```
# Train a Gradient Boosting Model

gbmodel1 = GradientBoostingRegressor(n_estimators = 1000,
                                     learning_rate = .005,
                                    max_depth = 4)
gbmodel1.fit(X_train,y_train)
```

Out[11]:
```
GradientBoostingRegressor(alpha=0.9, init=None, learning_rate=0.005,
             loss='ls', max_depth=4, max_features=None,
             max_leaf_nodes=None, min_samples_leaf=1, min_samples_split=2,
             min_weight_fraction_leaf=0.0, n_estimators=1000,
             presort='auto', random_state=None, subsample=1.0, verbose=0,
             warm_start=False)
```

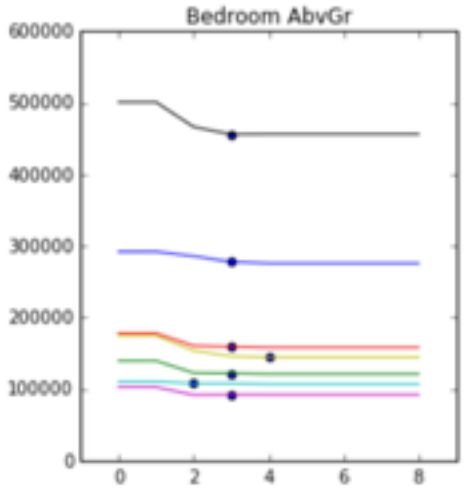
ML-Insights Example: Housing

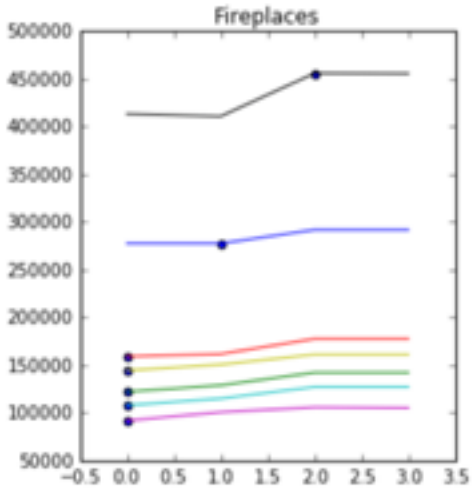
In [13]:
```
fig, ax = plt.subplots()

ind = np.array(range(len(X.columns)))+.7
plt.barh(ind,gbmodel1.feature_importances_);
ax.set_yticks(ind + .7);
ax.set_yticklabels((X_test.columns));
```

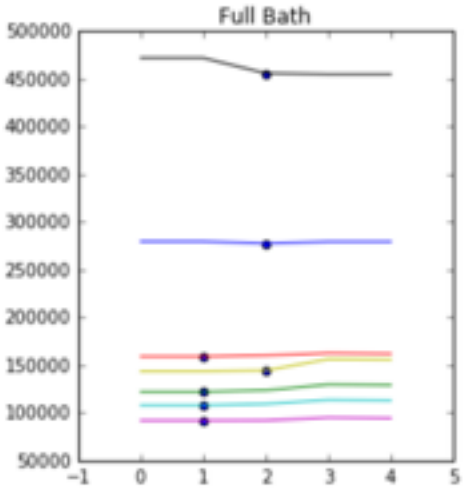


ML-Insights Example: Housing

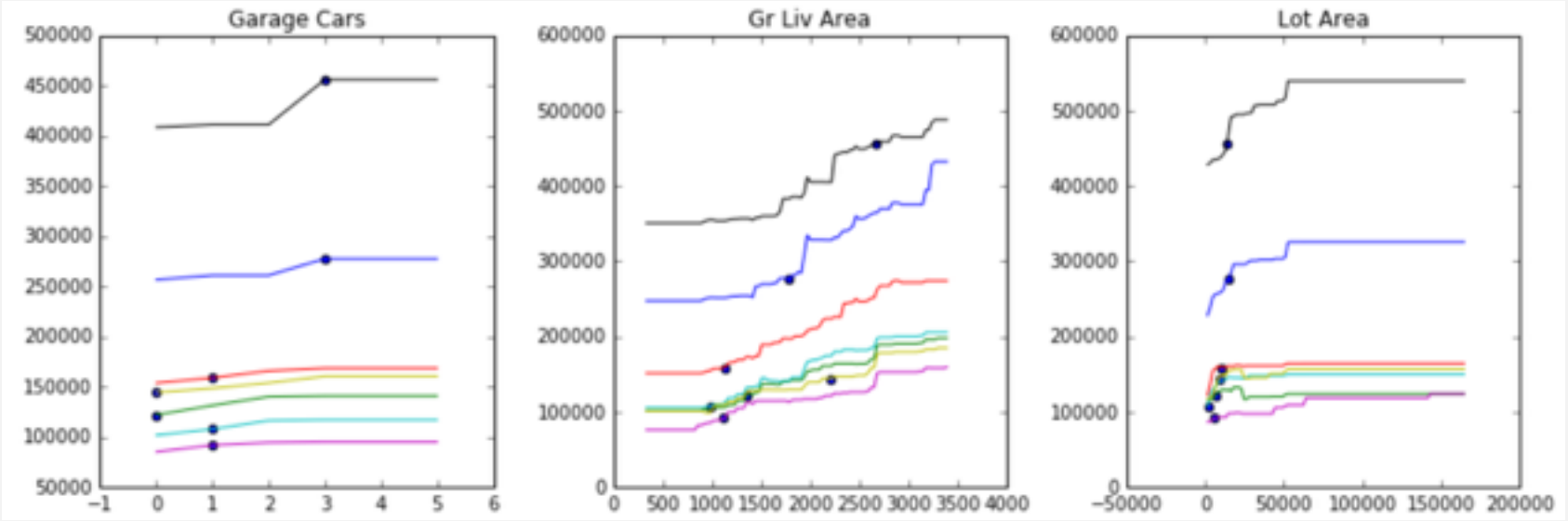
```
In [14]: mxr = mli.ModelXRay(gbmodel1,X_test)

In [15]: mxr.feature_dependence_plots(num_pts=7)
Out[15]: array([193, 300, 516, 861, 278, 782, 170])
```



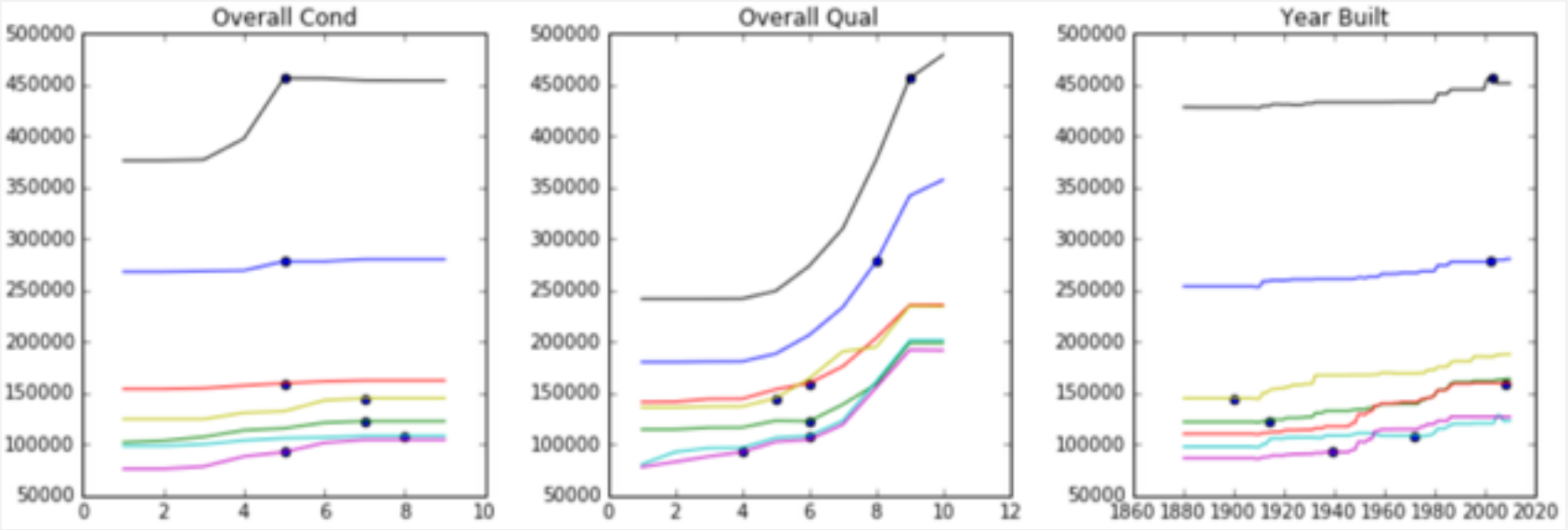




ML-Insights Example: Housing

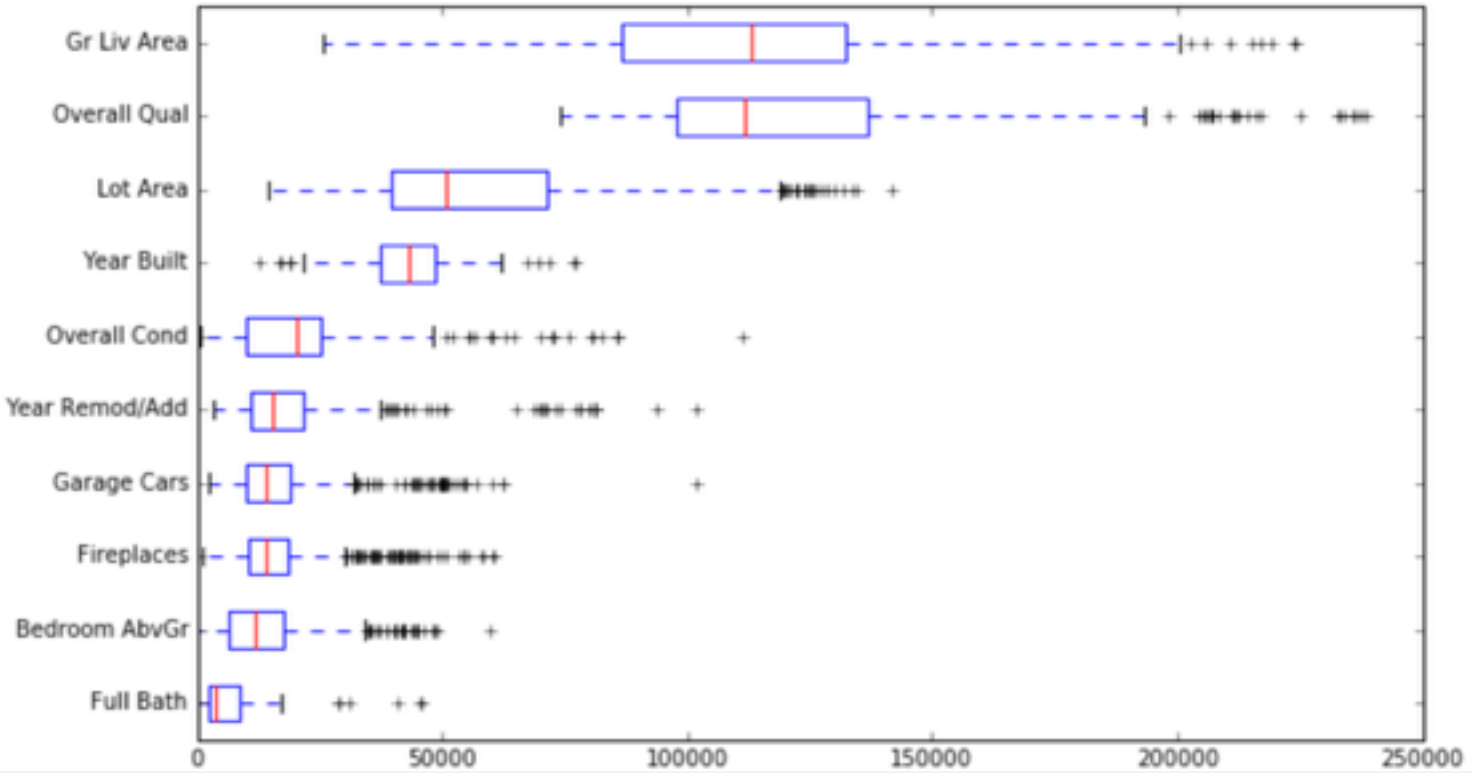


ML-Insights Example: Housing



ML-Insights Example: Housing

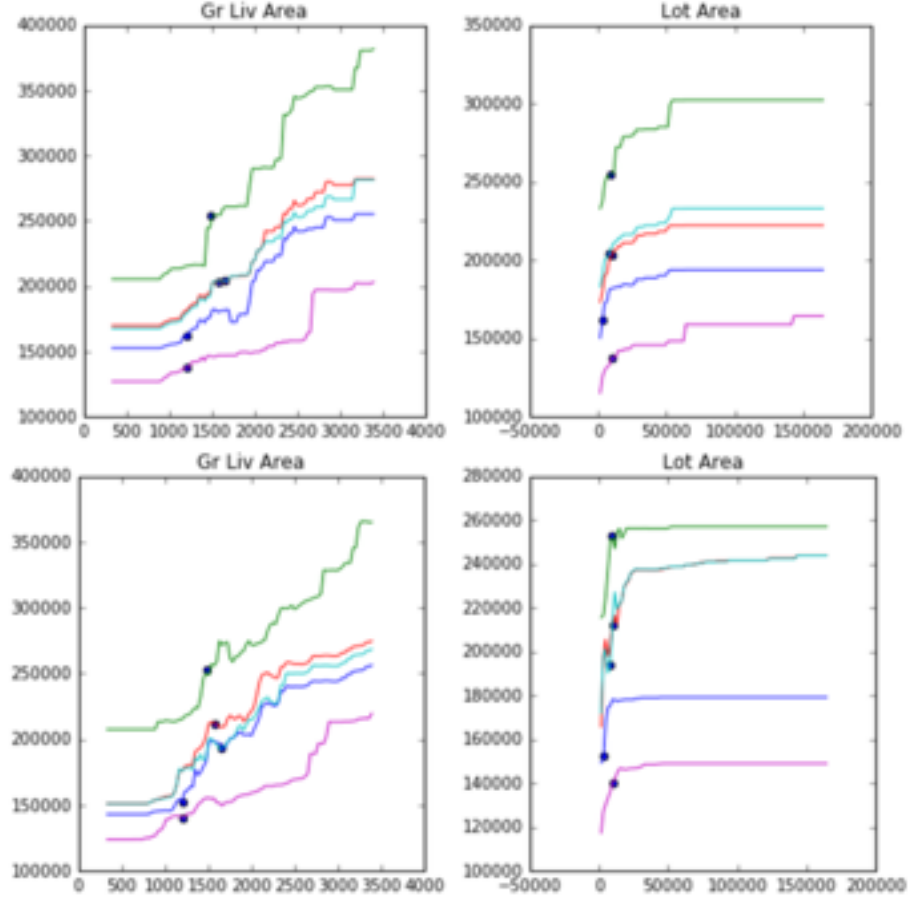
```
In [16]: mxr.feature_effect_summary()
```



ML-Insights Example: Housing

```
In [19]: diff_path_obj = mxr.explain_prediction_difference(193,300, tol=.05)

         Your initial point has a target value of 277564.7027
         Your final point has a target value of 122014.3857
         Changing Overall Qual from 8.0 to 6.0
                         changes your target by -71615.9773 to 205948.7254
         ----------
         Changing Year Built from 2002.0 to 1914.0
                         changes your target by -34047.5502 to 171901.1752
         ----------
         Changing Gr Liv Area from 1786.0 to 1355.0
                         changes your target by -22287.4459 to 149613.7293
         ----------
         Changing Garage Cars from 3.0 to 0.0
                         changes your target by -17095.9954 to 132517.7339
         ----------
         Changing Overall Cond from 5.0 to 7.0
                         changes your target by 11844.2895 to 144362.0233
         ----------
         Changing Lot Area from 14860.0 to 6882.0
                         changes your target by -16290.4091 to 128071.6142
         ----------
         Tolerance of 0.05 reached
         Current value of 128071.6142 is within 5.0% of 122014.3857
```

ML-Insights Example: Housing



Grad. Boosting

Random Forest

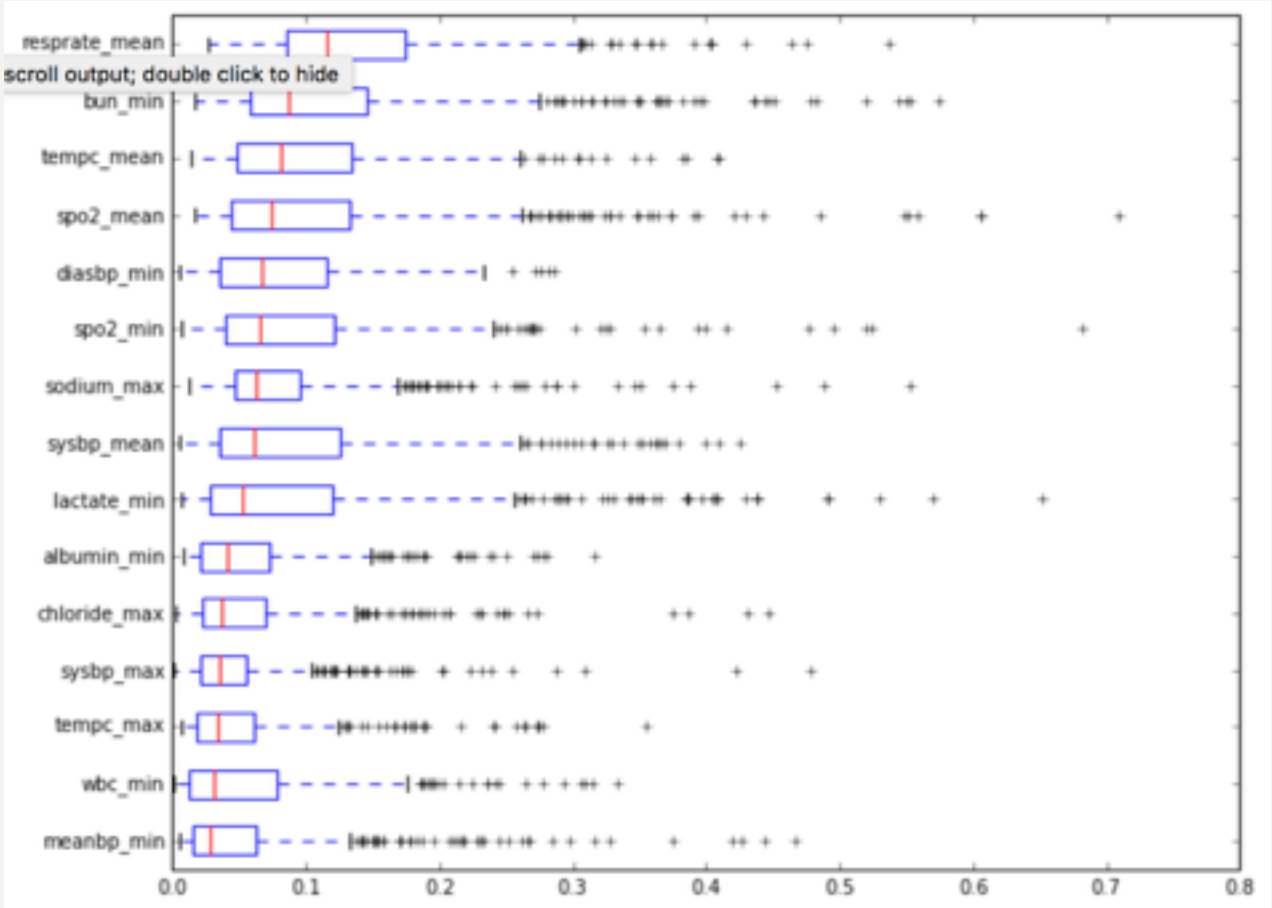
ML-Insights Example 2

Example 2: MIMIC Critical Care Database*

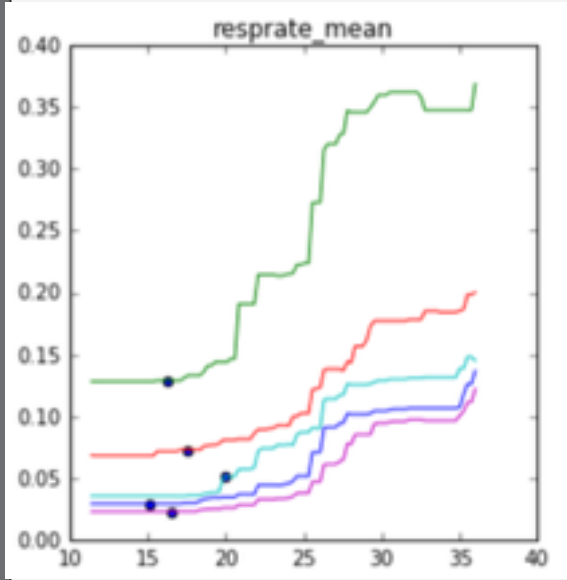
- Predict mortality in ICU Patients
- 59726 data points, split 70/30 train/test
- 51 labs / vitals aggregated over first 24 hours in ICU
- Fit three models: Gradient Boosting, Random Forest, and XGBoost

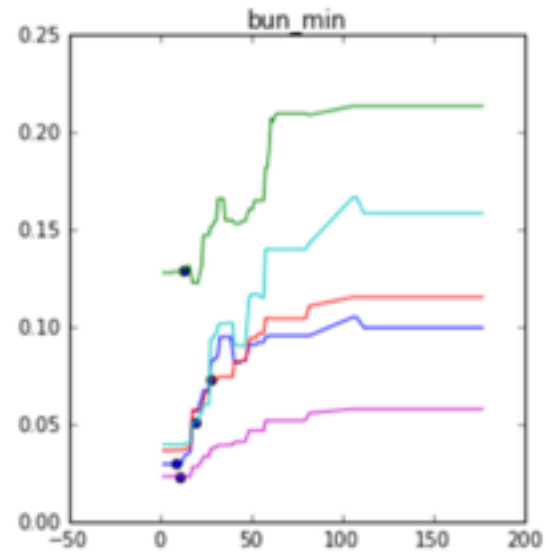
*MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016).
https://mimic.physionet.org

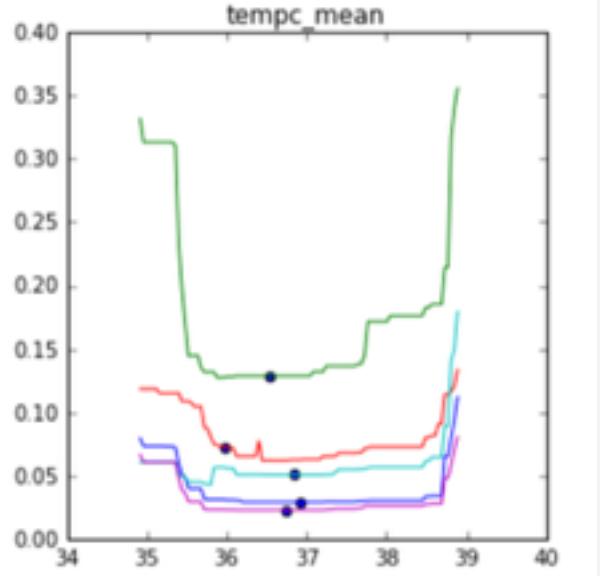
ML-Insights Package: Medical



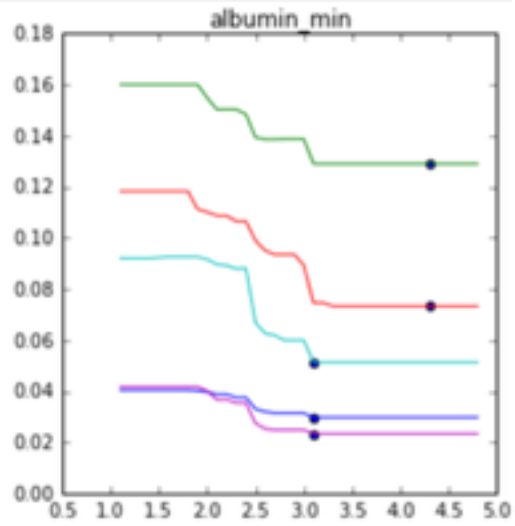
ML-Insights Package: Medical

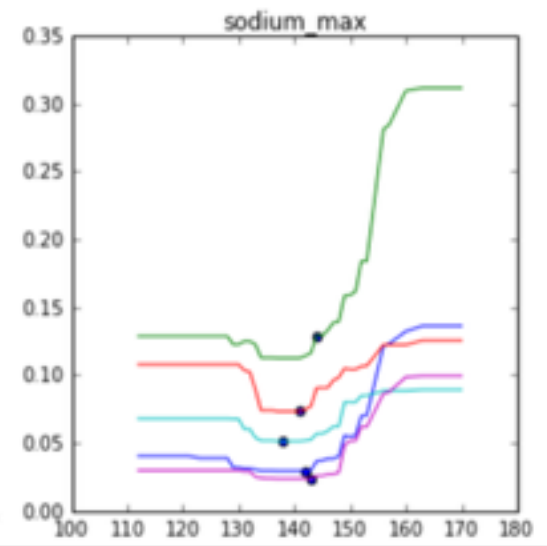


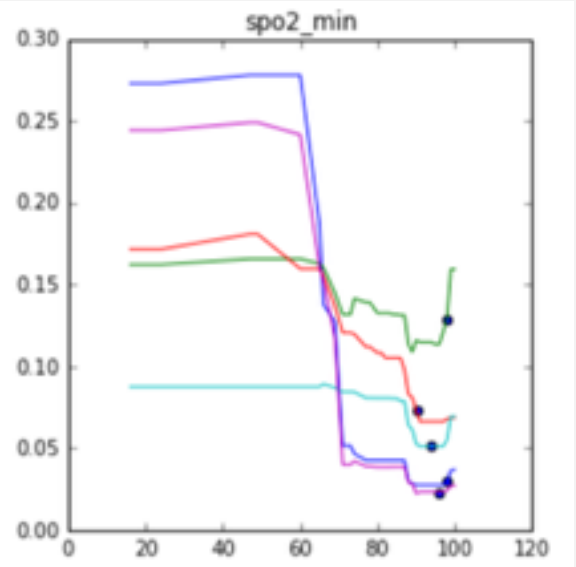




ML-Insights Package: Medical







ML-Insights Package: Medical

```
mxr.explain_prediction_difference(311,388,tol=.05)
```

```
Your initial point has a target value of 0.0255
Your final point has a target value of 0.1691
Changing sysbp_mean from 99.2143 to 79.2692
                changes your target by 0.031 to 0.0565
----------
Changing resprate_mean from 19.3333 to 25.6071
                changes your target by 0.0575 to 0.114
----------
Changing albumin_min from 3.1 to 2.4
                changes your target by 0.0593 to 0.1733
----------
Tolerance of 0.05 reached
Current value of 0.1733 is within 5.0% of 0.1691
```

Where to Find More

- To install: “pip install ml_insights”
- Github: https://github.com/numeristical/introspective
- Documentation: http://ml-insights.readthedocs.io
- Blog: www.numeristical.com
- Examples: https://github.com/numeristical/introspective/tree/master/examples