An Enriching Meal: Machine Learning

Nick Koenig, Ph.D., David Futrell, Ph.D., Matthew Arsenault, Ph.D.

Walmart

Daniel Schmerling, Ph.D.

Capital One

Walmart Labs

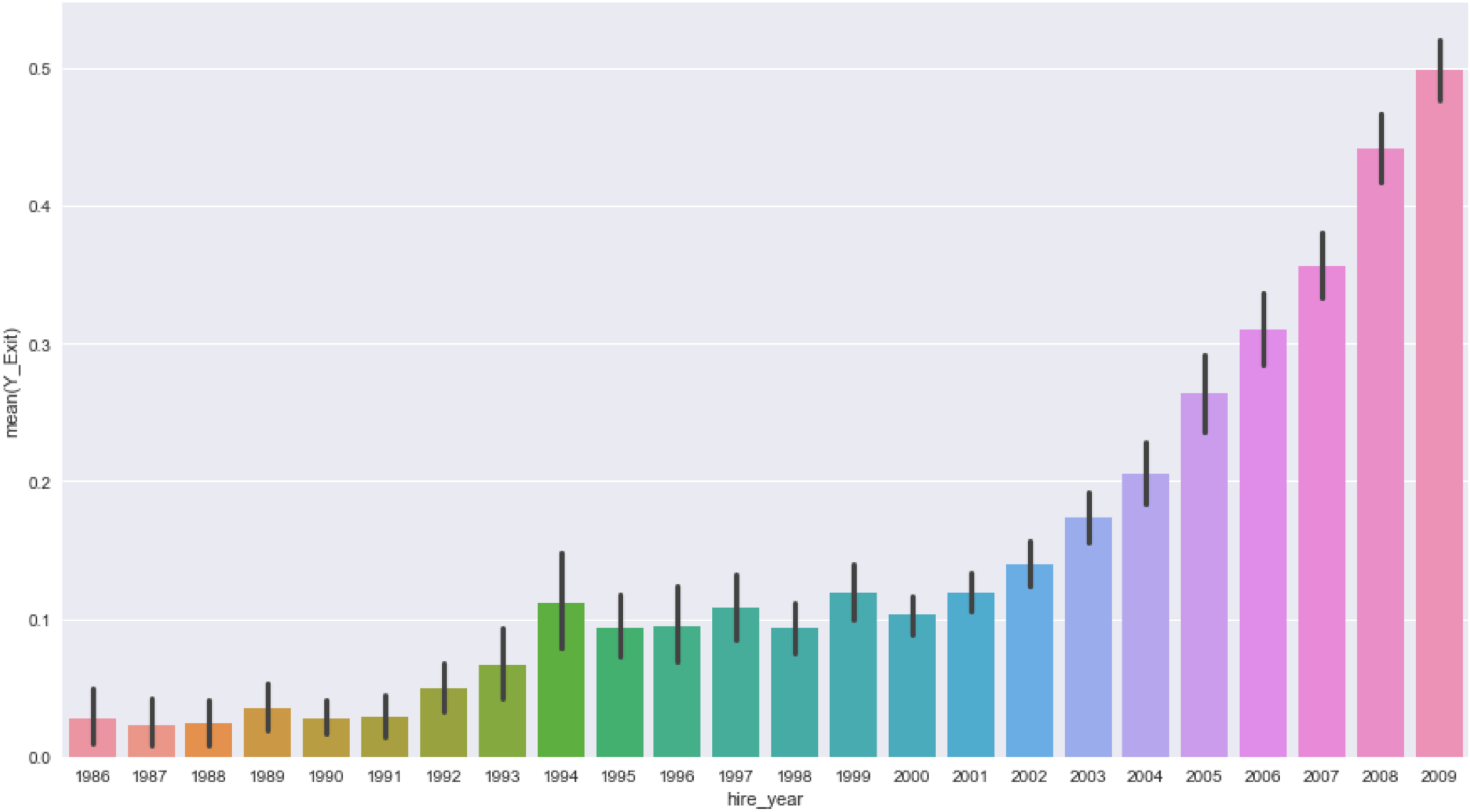
Agenda

Walmart Labs

General Approach

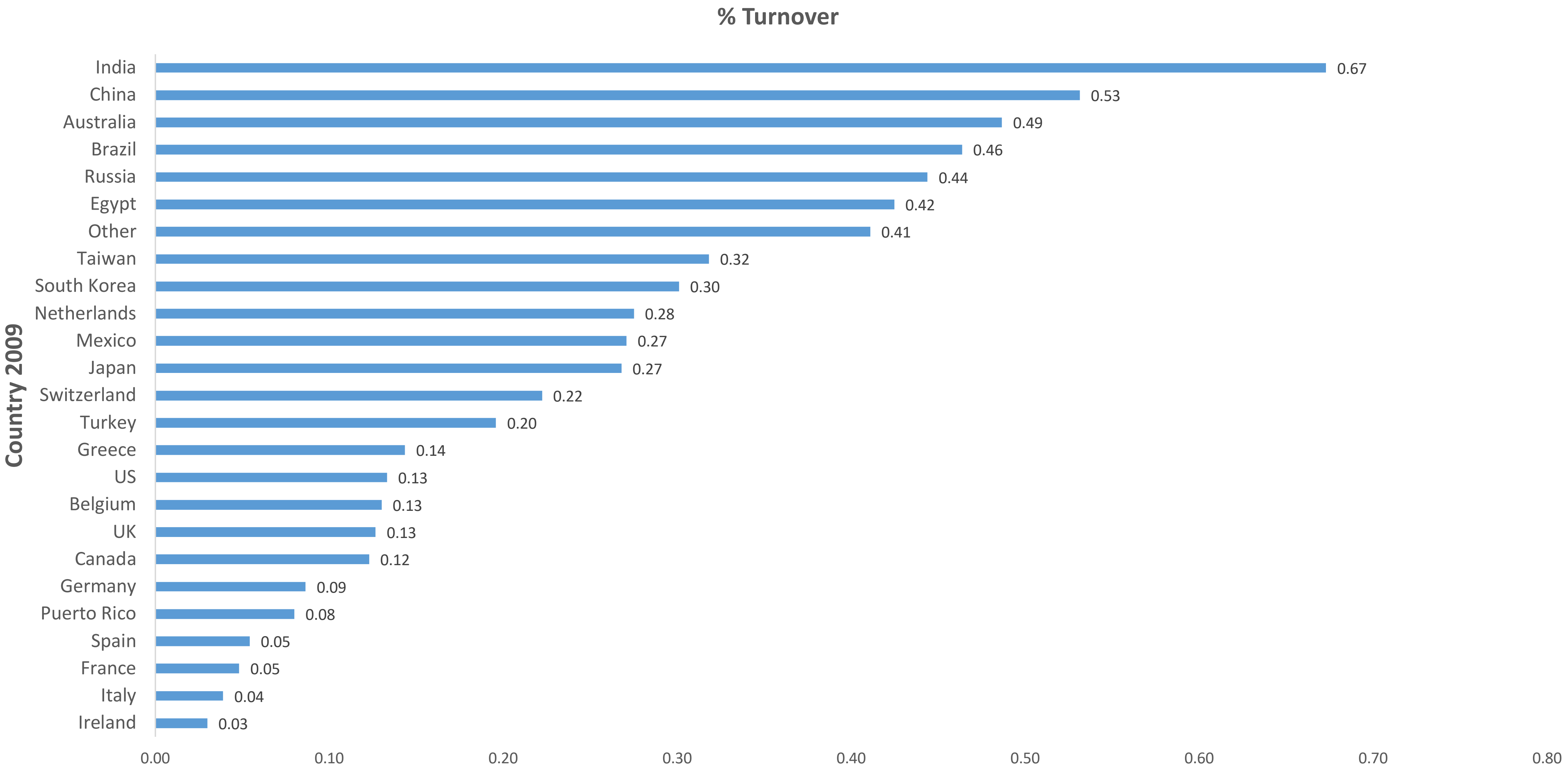
a) Reviewed the data structure and descriptive statistics (know your data)
b) Considered how the variables *should* relate to turnover (know your theory)
c) Explored results from correlations and logistic regression models
d) Engineered features to investigate potential non-linear effects (added noise)
e) Addressed missing data issues (non-existent vs. missing)
f) External sources of data (decided against inclusion)
g) Machine and deep learning algorithms (classification)
h) Explored algorithm enhancing procedures (boosting, blending, etc.)

EDA



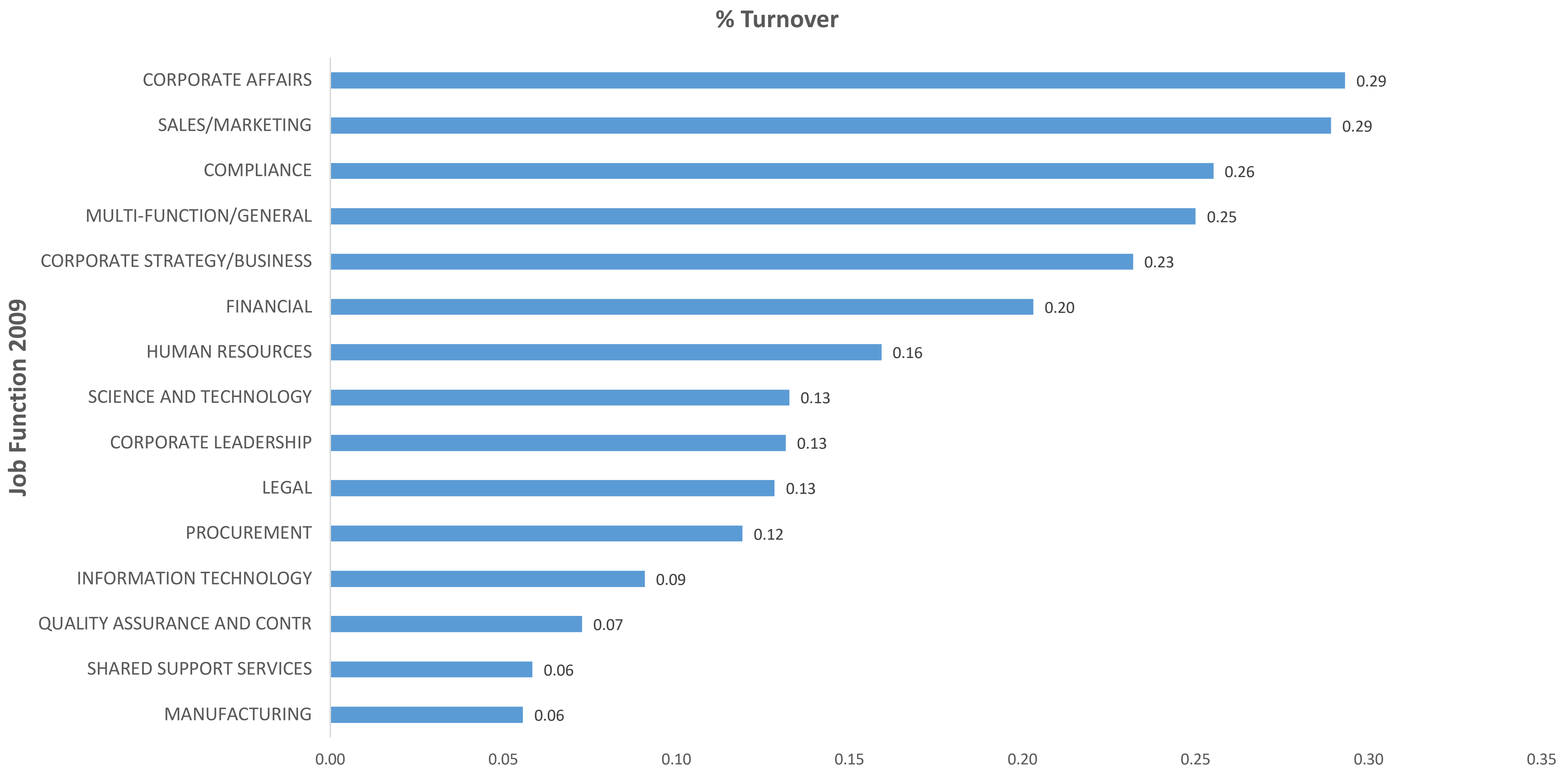
Walmart Labs

EDA



Walmart Labs

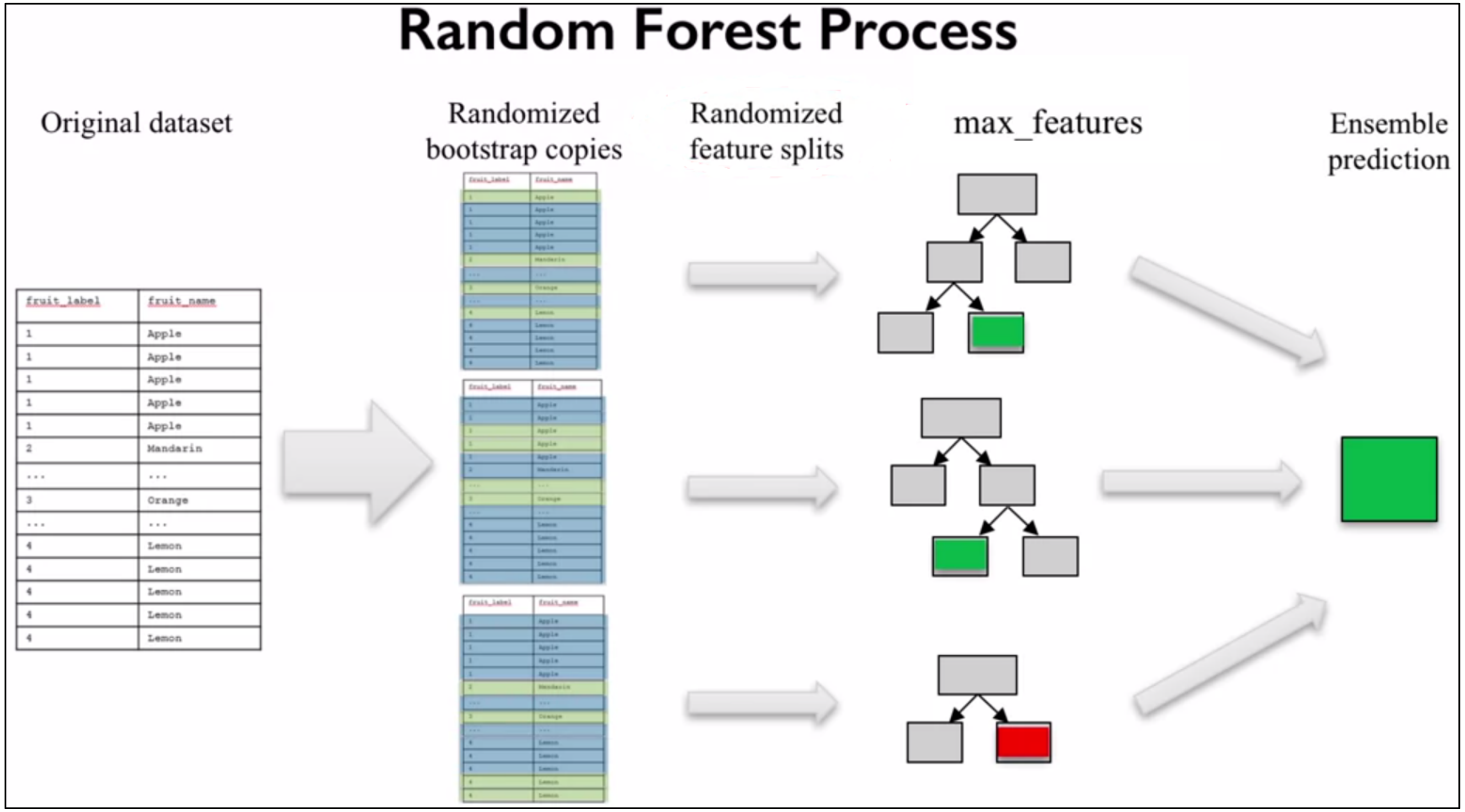
EDA



Walmart Labs

Random Forests & Bagging

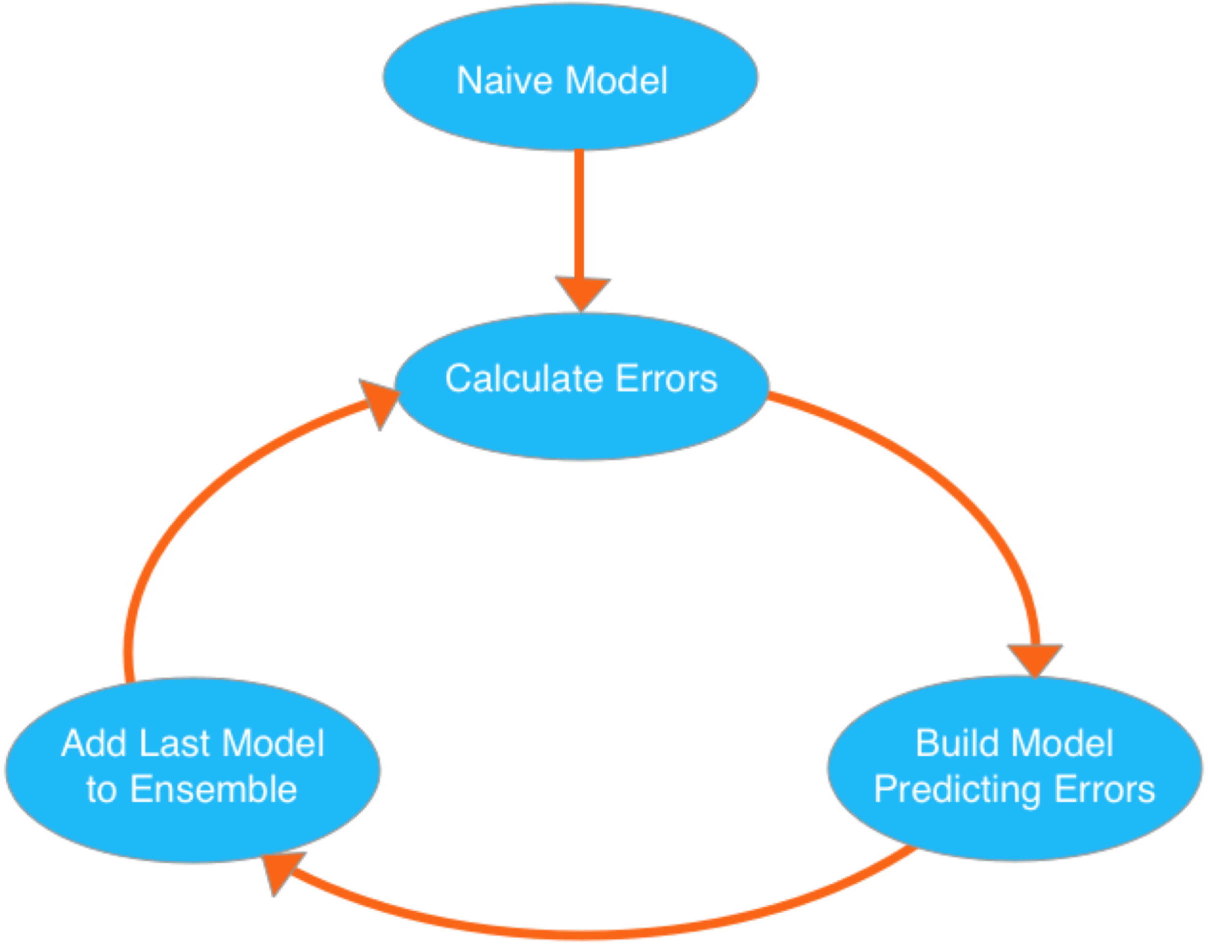
- ❑ Random Forests select random combinations of cases and features from the original dataset to create multiple decision trees
- ❑ Each decision tree is applied to a holdout dataset identifying the most likely outcome per tree
- ❑ The outcome most often identified for the case is used as the final prediction



Walmart Labs

Gradient Boosted Trees

- We used Boosting Ensemble Methods (XGBoost and LightGBM)
- The underlying idea is that we are taking a set of weak learners and using them to create a single strong learner

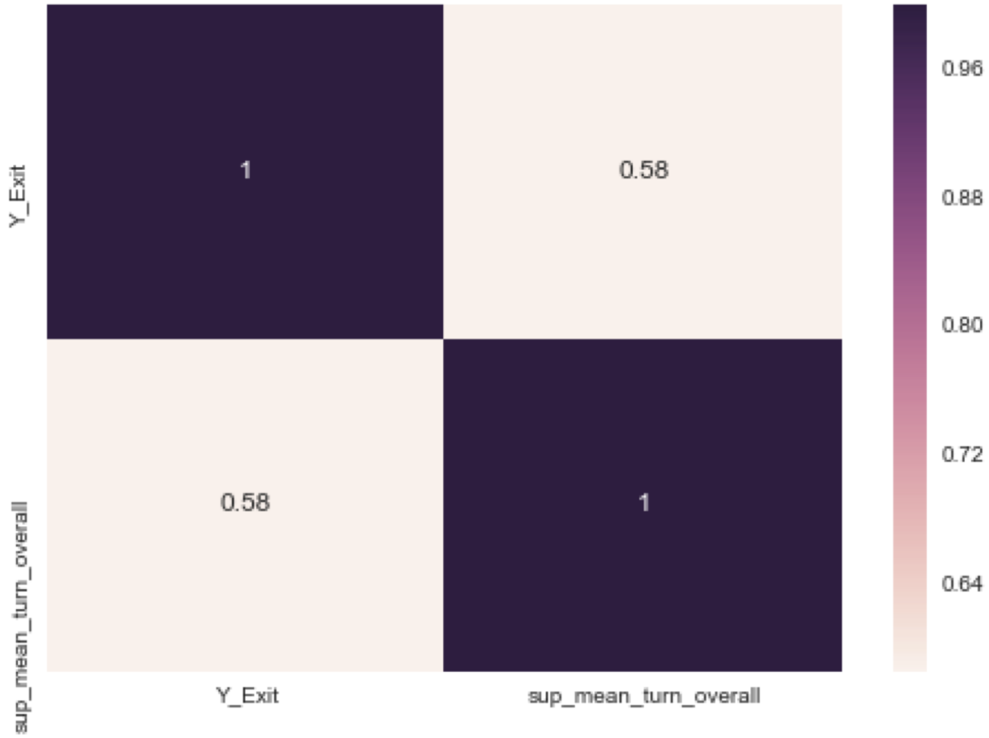


Deep Learning: Neural Networks

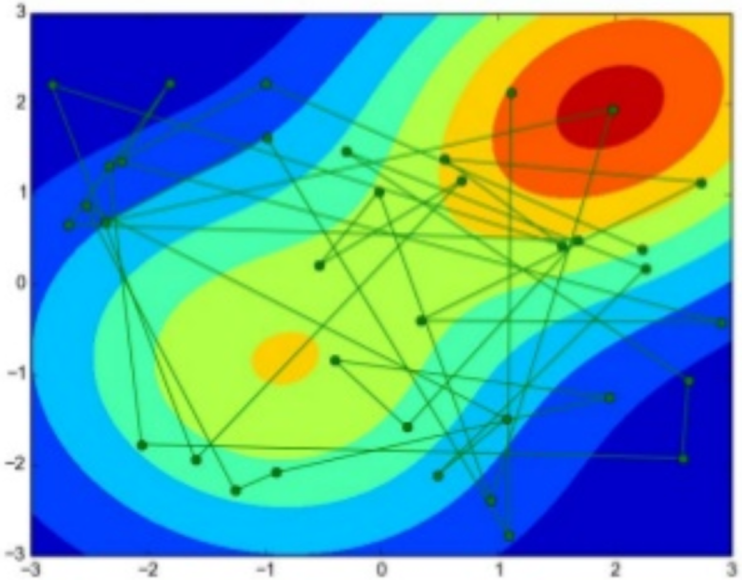
- Uses all features to determine the most likely outcome for each case
- Uses a training dataset to alter the feature weights until the difference in magnitude between the prediction and the actual outcome reaches a local minimum
- Once the local minimum is reached with the altered feature weights, this network of nodes connected by weights can be used on the holdout dataset to predict outcomes

Noise and Hyperparameter Turning

- Noise is necessary when target encoding high cardinality features and sometimes it just doesn't work



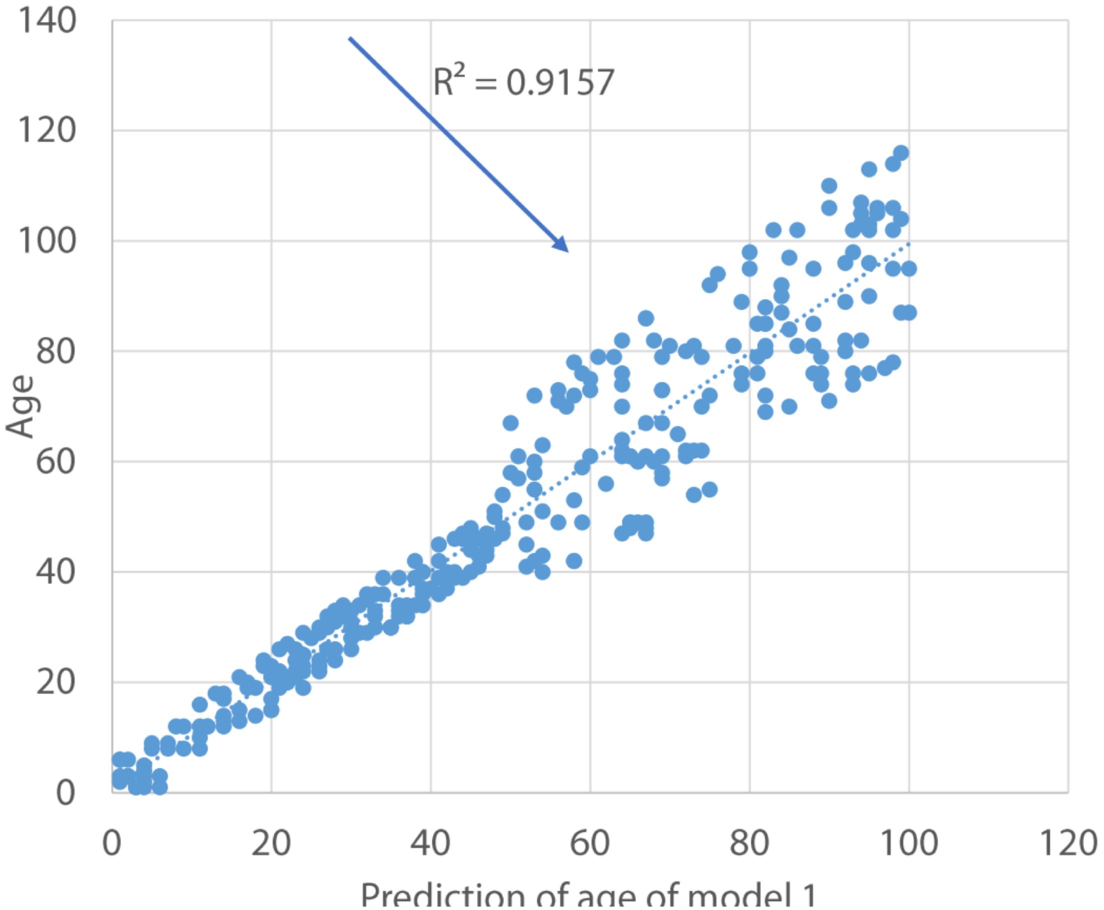
- Hyperparamaters are *variables* that cannot be appropriately estimated by the data itself (k in k-means clustering)



```
param_dist = {"learning_rate":np.linspace(.009,.20,50),
              "max_depth": [-1,3,5,7,10,12,15],
              'min_child_samples': sp_randint(2,20),
              "reg_lambda":[0.0,0.01,0.1],
              "n_estimators": [100,250,500,750,1000]}
```

Blending

Averaging ensemble methods to improve variance explained



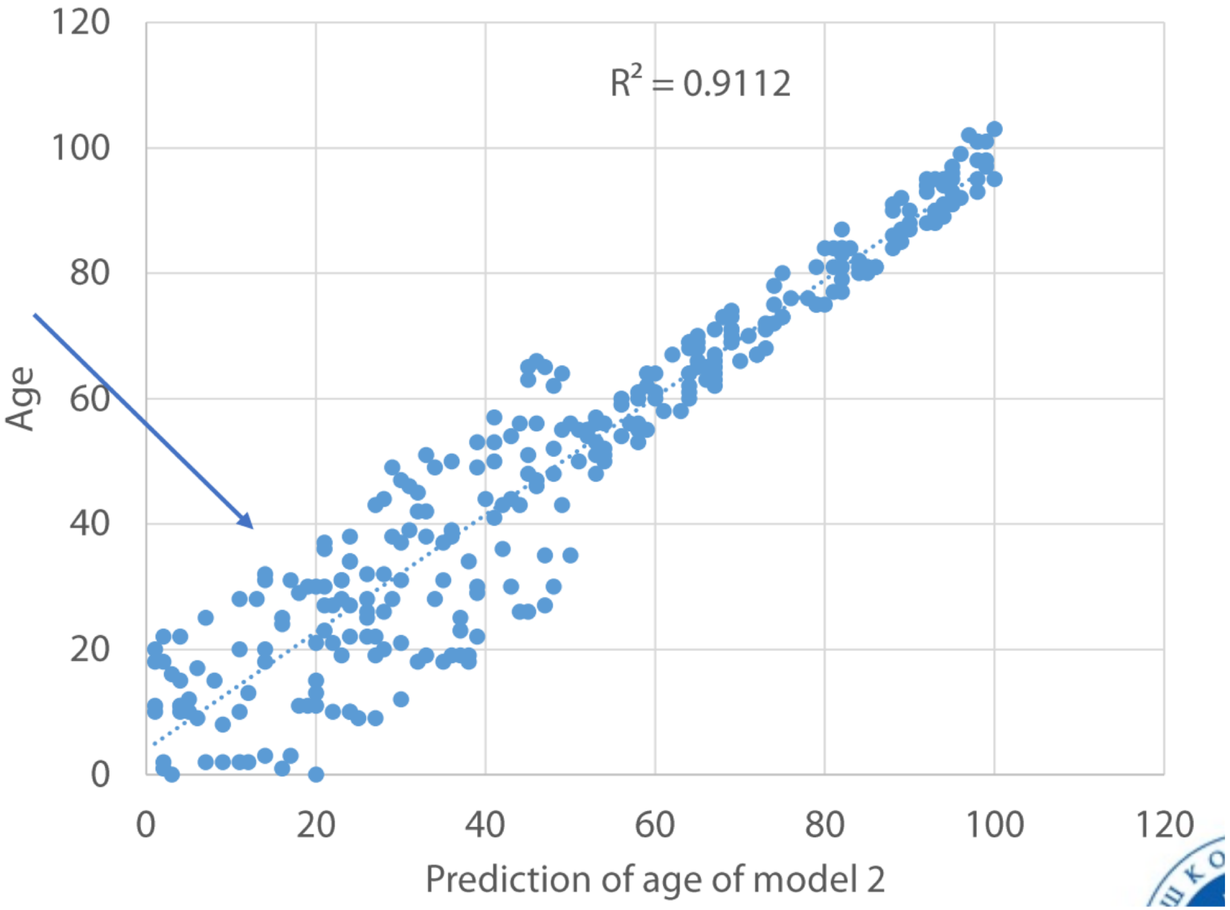


Image Source: https://www.coursera.org/learn/competitive-data-science/lecture/MJKCi/introduction-into-ensemble-methods

Blending

Averaging ensemble methods to improve variance explained

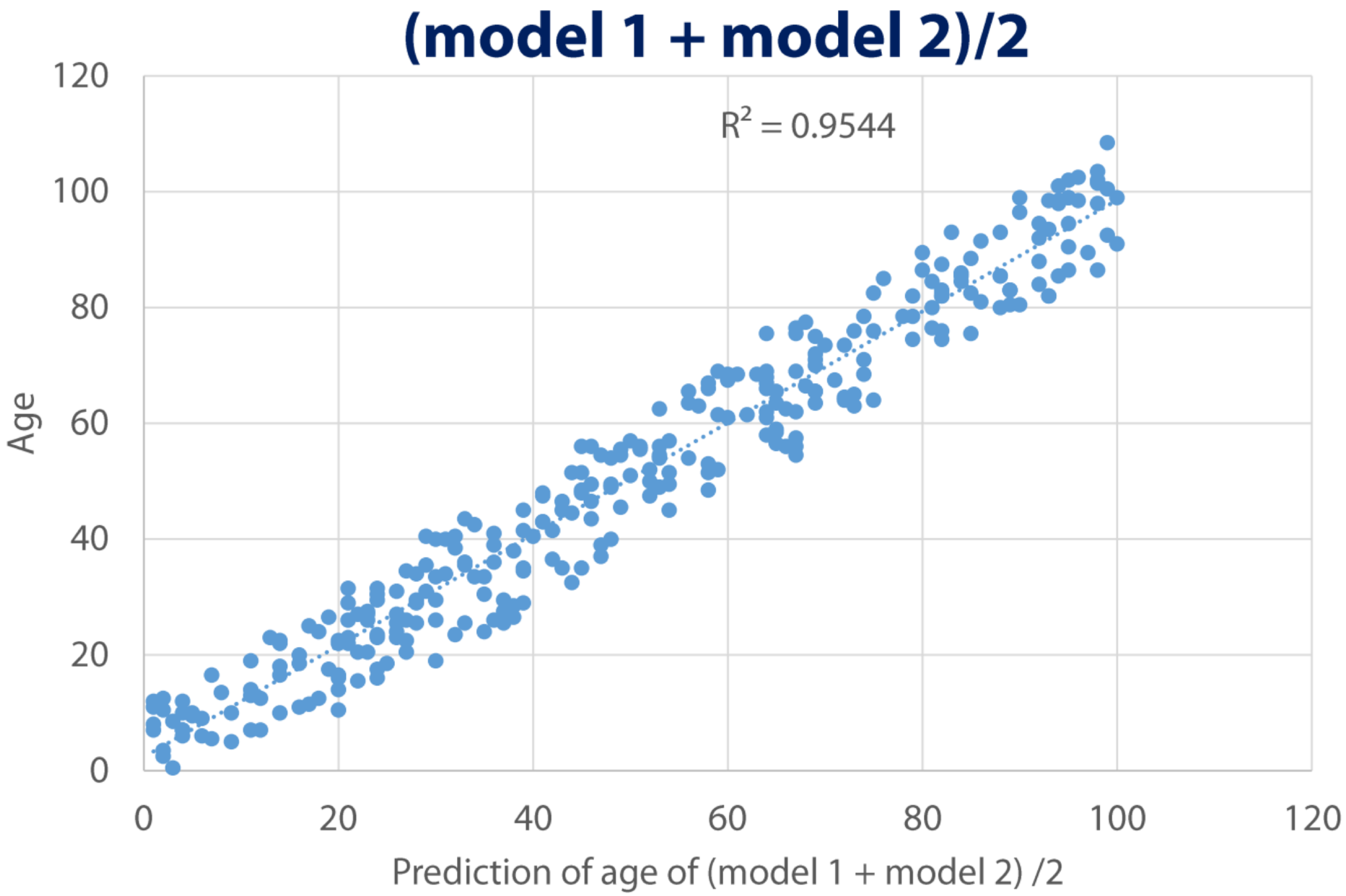
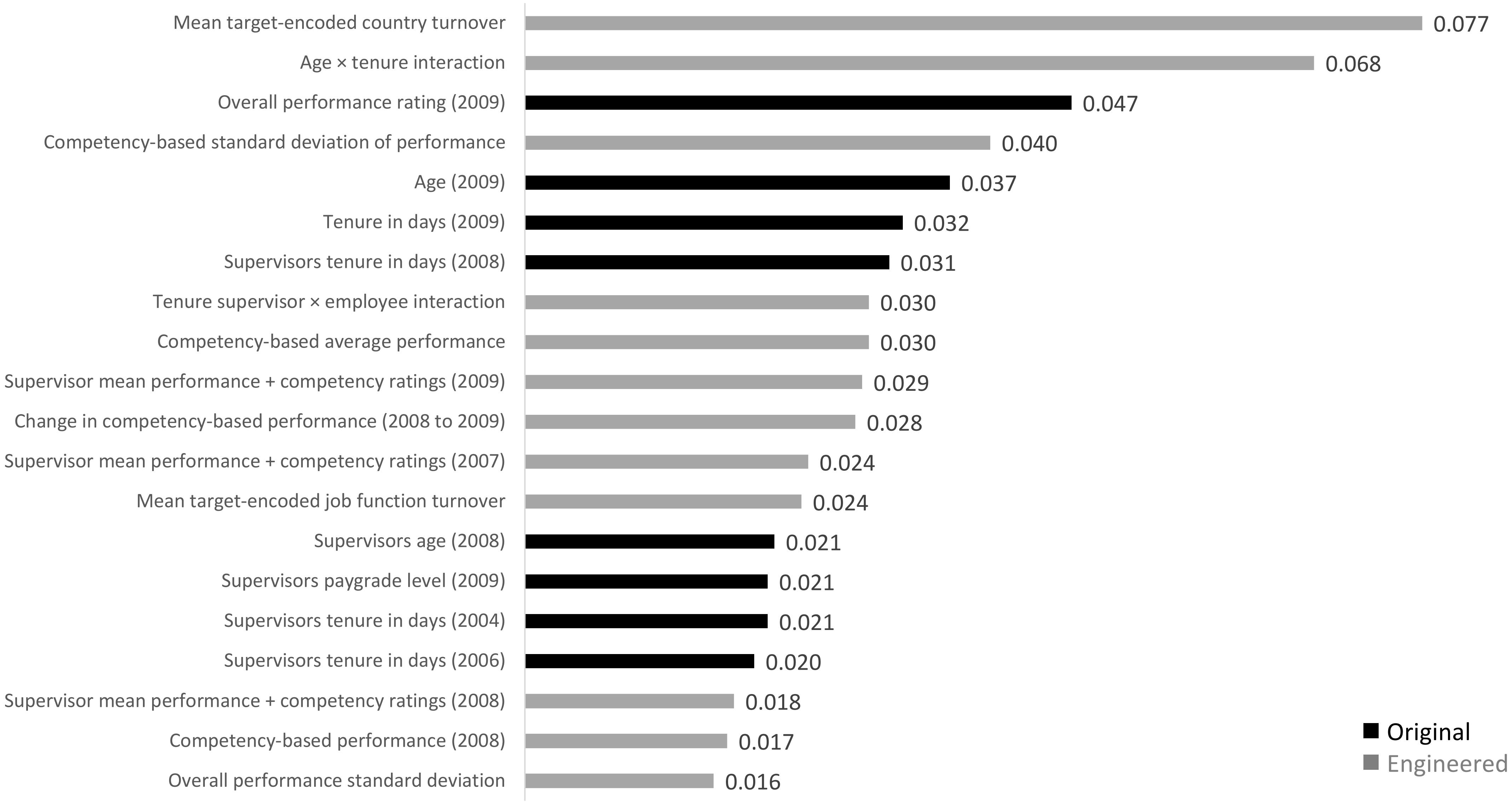


Image Source: https://www.coursera.org/learn/competitive-data-science/lecture/MJKCi/introduction-into-ensemble-methods

Walmart Labs

20 most Import Features in Best Performing Algorithm




```
In [37]: stacked_test_df['Y_ExitProbability'] = ((stacked_test_df['XGB']*.35)+(stacked_test_df['LGBM']*.3)+
                                                 (stacked_test_df['RF']*.05)+
                                                 (stacked_test_df['ANN']*.3))
```

Walmart Labs

Conclusions

- ❑ Machine learning methods can improve prediction; but these methods are not a substitute for good science
  - Prediction is not understanding
- ❑ Certain practices should be applied more broadly to increase confidence
  - Train/test splits should be standard practice in model testing
- ❑ Underlying algorithm assumptions need to be discussed/explored
  - How are cases with missing data treated?
  - How are features with high multicollinearity treated?
- ❑ Blending theory, measurement, and data science methods offers exciting opportunities for both research and practice

I can calculate the motion of heavenly bodies, but not the madness of people

Isaac Newton