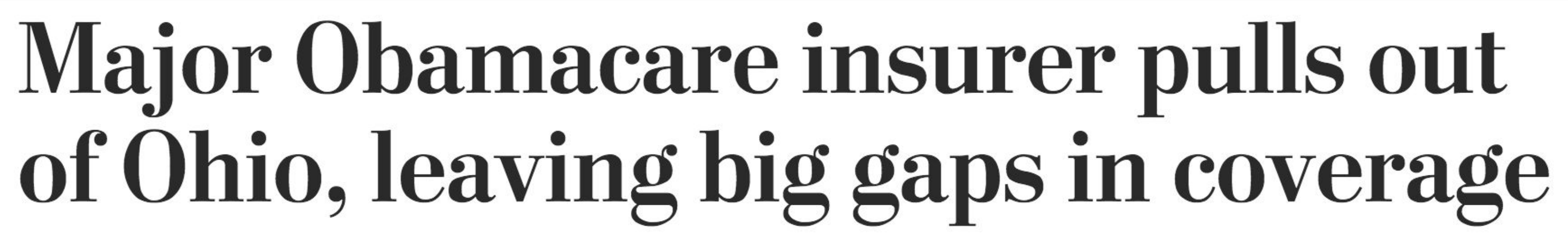
Project McNulty

Predicting Whether Health Insurers Will Exit from Local Markets

Early Warning System for Non-Group Individual Insurance Market



Major Obamacare insurer pulls out of Ohio, leaving big gaps in coverage

“Health insurer Anthem is pulling out of Ohio's Affordable Care Act marketplace, a move that leaves people in a fifth of the state's counties facing the prospect of having zero insurers selling individual marketplace plans in their area next year.”

Small Data Set and Well Balanced Binary Classes

23,000 total county-level observations for years 2015, 2016, 2017 combined

<u>MESSY SOURCE DATA</u>

Engineered Target and about 10 features / 5 included in main models

<u>BINARY CLASSIFICATION</u>

65% of observations have insurer offering insurance in county next year
35% of observations have insurer leaving county and not offering insurance

F1 Score and Baseline Models

Focused on F1 Score
Want to warn before exit occurs

Too many false alarms -> No action

Balance precision and recall

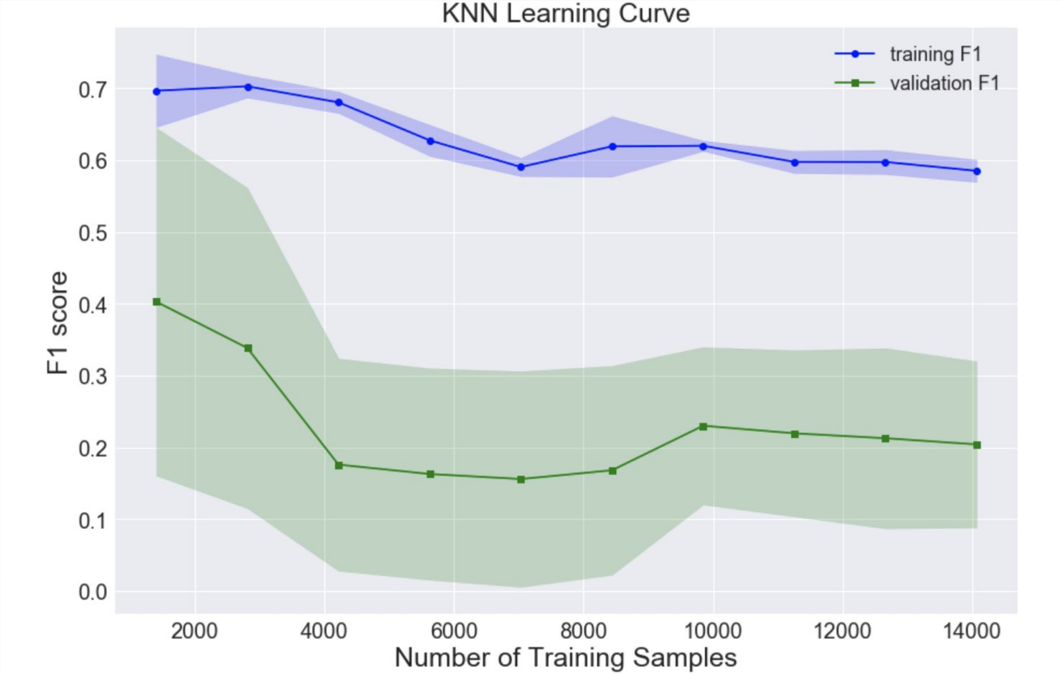
Models	**Validation / Selection** **F1 Scores**
Logistic Regression	0.005
Naive Bayes	0.095
KNN	0.20

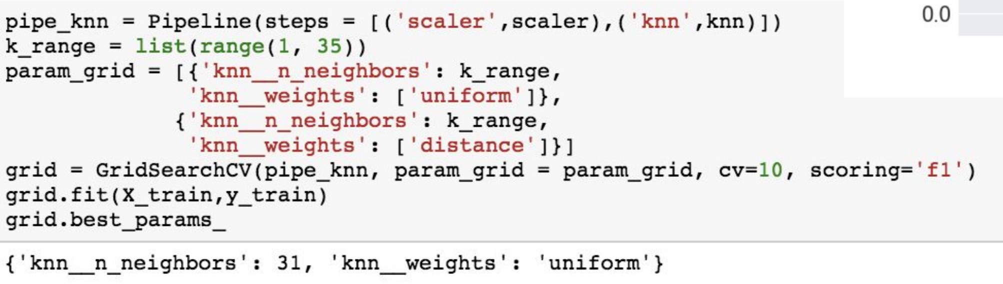
Closer Look at KNN to Diagnose What May Improve Performance

Grid search on KNN

Learning curve with hyperparemeter values from grid search

Look at Random Forest to see if ensemble method can reduce variance





```
pipe_knn = Pipeline(steps = [('scaler',scaler),('knn',knn)])
k_range = list(range(1, 35))
param_grid = [{'knn__n_neighbors': k_range,
              'knn__weights': ['uniform']},
             {'knn__n_neighbors': k_range,
              'knn__weights': ['distance']}]
grid = GridSearchCV(pipe_knn, param_grid = param_grid, cv=10, scoring='f1')
grid.fit(X_train,y_train)
grid.best_params_
```

```
{'knn__n_neighbors': 31, 'knn__weights': 'uniform'}
```

Random Forest and Holdout Test

Performed grid search

F1 of 0.33 for Random Forest versus 0.20 in KNN in validation sets

Full 20% hold out set

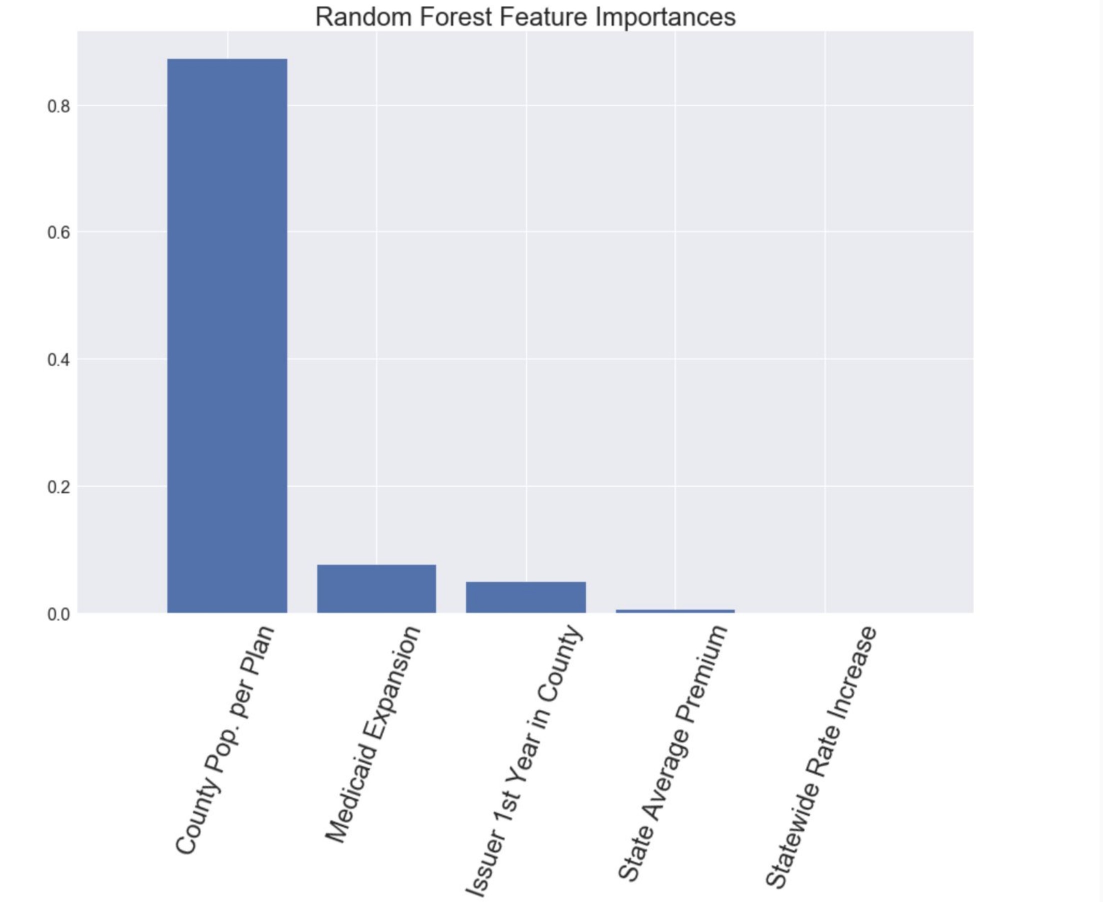
Based on Insurance Issuers not appearing in any previous training or validation sets

Random Forest on Holdout Set

Accuracy:	0.51
Precision:	0.43
Recall:	0.11
F1:	0.18

Highest Feature Importance

County Population divided by Total Plans in County



Conclusions and Extensions

Not showing high enough F1 to be meaningful for user
More features would be priority

Possible Extensions to Improve Features
Compositional effects in plans offered
Overall rates may increase but plan level benefits more generous
Try to make rate pricing apples to apples for benefit changes

Is county split right?
Had to use state level data for some features for county level target
Spending time on edge cases of when insurer leaves some but not all counties in a state

More Data
Incorporate 12 additional states not in this data system

Appendix

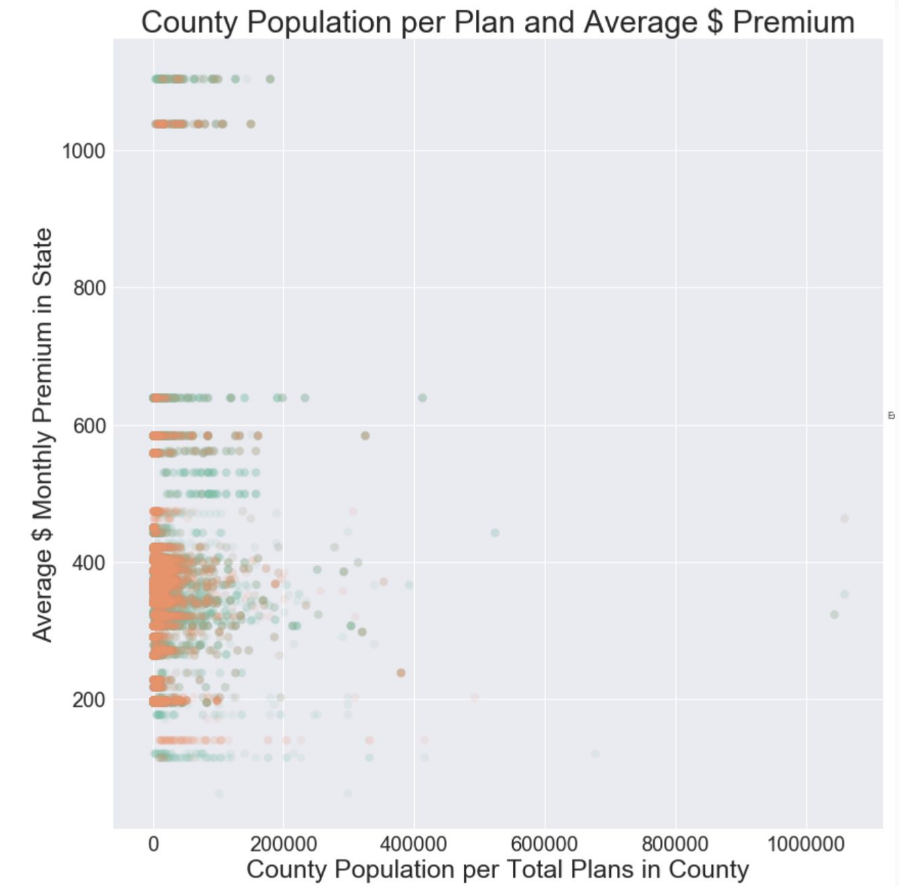
Feature Examples and EDA

EDA Suggests Complex Model

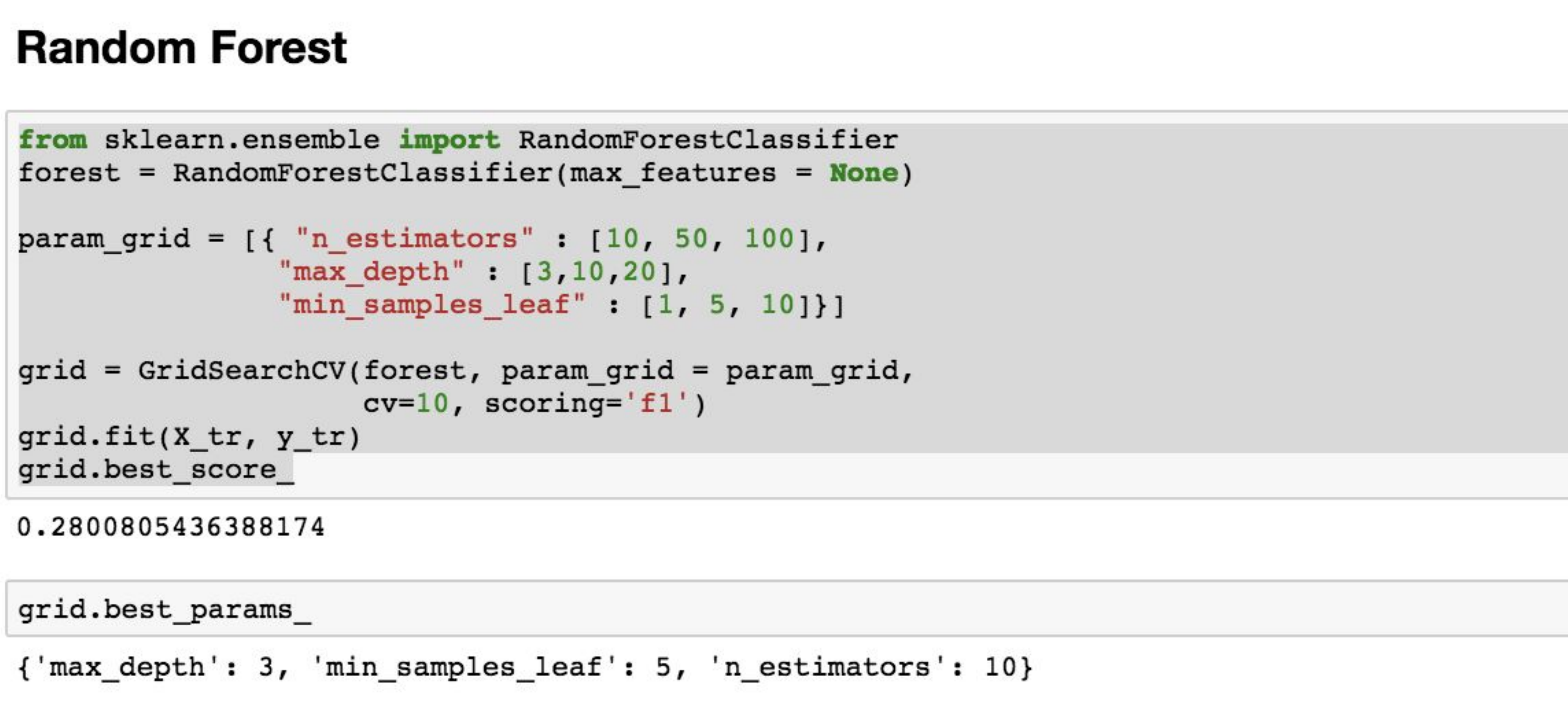
X-Axis
County Population divided by Number of Plans Offered in County

Y-Axis
Average $ Monthly Insurance Premium in State

Shading / Target Classifications
Green:	Offer Plan in County Next Year
Orange:	Leave County Next Year



Random Forest Grid Search



Random Forest

```
from sklearn.ensemble import RandomForestClassifier
forest = RandomForestClassifier(max_features = None)

param_grid = [{ "n_estimators" : [10, 50, 100],
               "max_depth" : [3,10,20],
               "min_samples_leaf" : [1, 5, 10]}]

grid = GridSearchCV(forest, param_grid = param_grid,
                    cv=10, scoring='f1')
grid.fit(X_tr, y_tr)
grid.best_score_
```

```
0.2800805436388174
```

```
grid.best_params_
```

```
{'max_depth': 3, 'min_samples_leaf': 5, 'n_estimators': 10}
```