Civilization 6 (Rise & Fall) start city settlement analysis

Leon Kotze, December 2018

Executive Summary

The document presents an analysis of data concerning Civilization 6 (Rise & Fall) start city settlement location and eventual prosperity. The data was generated by the author with the intention to use a Machine Learning model to predict city prosperity based solely on tile (aka plot) information available at the start of the game.

Prosperity is defined as the combined decile score across 5 key yields. The top quartile (top 25%) of these combined scores were deemed as good, aka prosperous cities. The remaining 75% were deemed poor. For example, a city is in the top decile of Food, Production, Gold. It is also in the 3rd decile of Science and Culture. This translates to a score of 41 (9*3 + 7*2) and would put it in the top quartile, i.e. a good city.

The input features consist of the terrain/feature and resource counts of the tile the city is settled on, and the surrounding 18 tiles within 2 tiles distance. The counts are converted into percentage of the city total. That is, the counts are all divided by 19. I also included whether the city is next to a river, or not.

The significant input features were:

- *GrasslandHillsWoods* - Percentage of tiles that are Grassland with Hills and Woods.
- *Lux* - Percentage of tiles that have luxury resources.
- *Stone* - Percentage of tiles that have Stone as a resource.
- *Rice* - Percentage of tiles with Rice bonus resource.
- *Bananas* - Percentage of tiles with Bananas bonus resource.
- *Wheat* - Percentage of tiles with Wheat bonus resource.
- *Deer* - Percentage of tiles with Deer bonus resource.
- *Fish* - Percentage of tiles with Fish bonus resource.
- *PlainsWoods* - Percentage of tiles that are Plains with Woods.
- *PlainsHilllsWoods* - Percentage of tiles that are Plains with Hills and Woods.
- *GrasslandWoods* - Percentage of tiles that are Grassland with Woods.
- *PlainsRainforest* - Percentage of tiles that are Plains with Rainforest.
- *Grassland* - Percentage of tiles that are Grassland.
- *GrasslandHills* - Percentage of tiles that are Grassland with Hills.
- *CoastLake* - Percentage of tiles that are Coast and Lake.
- *Plains* - Percentage of tiles that are Plains.
- *GrasslandMountain* – Percentage of tiles that are Grassland with a Mountain
- *cityHasRiver* - 0 if the city doesn't have a river, 1 if the city does.

Initial Data Exploration

As I am also the creator of the dataset, I will also cover the data creation process.

Input feature creation and analysis

The input used for the classification model consists of the city centre tile and surrounding 18 tiles within 2 tiles from said city centre. That is, for each of the 502 cities captured I captured 19 tiles worth of Terrain, Feature, and Resource counts. I wanted to create categories (and counts in categories) for each of the rings (city centre – 0, adjacent to city centre -1, and those two tiles away – 2). This would have produced in excess of 7K features which would need a lot more data than I was willing to capture.

I settled for a compromise where I combined Terrain and Feature into one category type. I also took the resources and grouped many together into OtherBonus as category type. I also made no distinction between the distance from the city centre. This reduced the input to 36 features which is a more manageable number.

The 9538 (19 tiles x 502 cities) plots has the following statistics:

Terrain	% of total
Grassland	26.0%
Plains	25.3%
Plains (Hills)	10.4%
Coast and Lake	10.4%
Grassland (Hills)	10.2%
Desert	5.6%
Rest	12.1%

Feature	% of total
None	73.0%
Woods	13.6%
Rainforest	8.6%
Marsh	2.4%
Floodplains	1.7%
Reef	0.6%
Oasis	0.1%

Resource	% of total
None	78.4%
Stone	2.9%
Wheat	2.3%
Cattle	1.3%
Rice	1.2%
Bananas	1.1%
Sheep	1.0%

Rest (each less than)	1.0%

Terrain & Feature	% of total
Grassland	18.3%
Plains	14.8%
Coast and Lake	9.8%
Grassland (Hills)	8.0%
Plains (Hills)	6.8%
PlainsRainforest	6.3%
GrasslandWoods	5.3%
PlainsWoods	4.1%
Desert	3.8%
Grassland Mountain	3.0%
Grassland Marsh	2.4%
DesertHills	2.3%
Plains (Hills) Rainforest	2.3%
Plains Mountain	2.2%
Grassland (Hills) Woods	2.2%
Desert Floodplains	1.7%
Tundra	1.6%
Plains (Hills) Woods	1.4%
Ocean	1.0%
Desert Mountain	0.7%
Coast and Lake Reef	0.6%
Tundra (Hills)	0.5%
Tundra Woods	0.4%
Tundra Mountain	0.2%
Desert Oasis	0.1%
Tundra (Hills) Woods	0.1%
Snow	0.0%

Of the 9538 tiles, 3663 (38.4%) has a river adjacent to it. If we look at the city centres only, of the 502 cities 410 (81.7%) were settled next to a river.

Output label creation and analysis

In order to predict a prosperous aka good city you first need to define what a good city is. I tried a few methods and, in the end, settled on using quantiles to rank and grade the cities. For each of the core yields; Food, Production, Gold, Science, and Culture I created a cumulative count over the 50 turns since the city was settled. Then at turn 50 I took this cumulative yield, broke it into deciles, and allocated a score of 0 through 9 from lowest to highest decile.

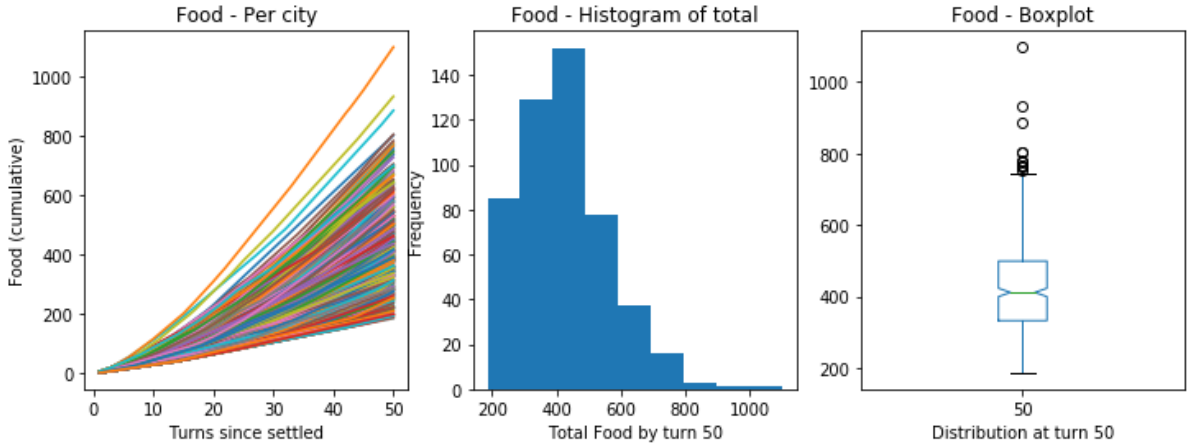
To calculate the cityTotal I summed these 5 numbers. The reasoning being that a good city would be in the upper deciles across the board. Good cities are those deemed to be in the top quartile (25%) of this combined score and were marked with a 1.

Next, we can look at each of the yields in isolation.

Food

The cumulative Food produced over 50 turns ranged between 186.00 and 1098.00 with an average of 423.57. If you think about the maximum number, the highest yielding food city was averaging nearly 22 Food per turn. This is almost unbelievable territory.

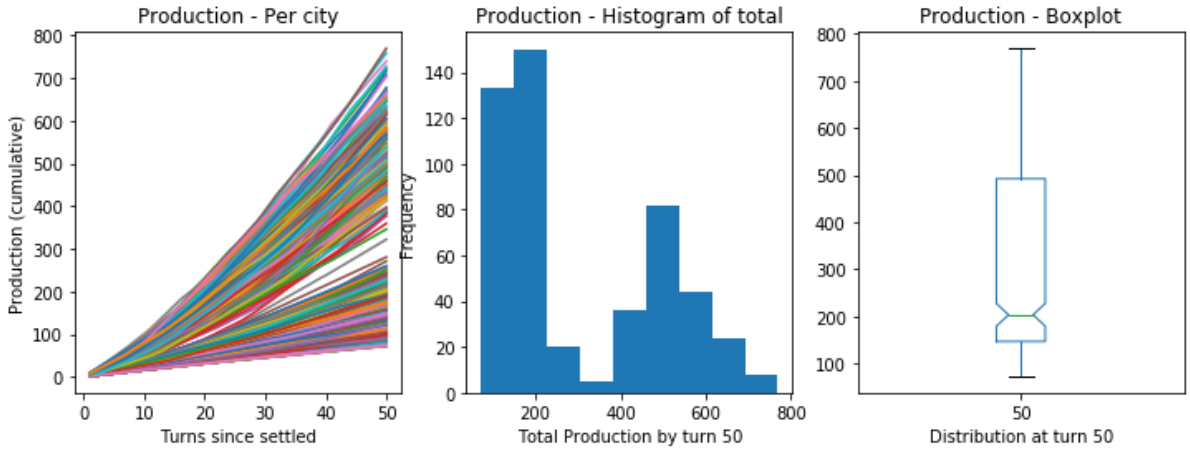
The visual representation of the Food produced looks as follows:



Production

The cumulative Production produced over 50 turns ranged between 72.30 and 768.80 with an average of 303.53. The median is 202.88 and the histogram is clearly bi-modal.

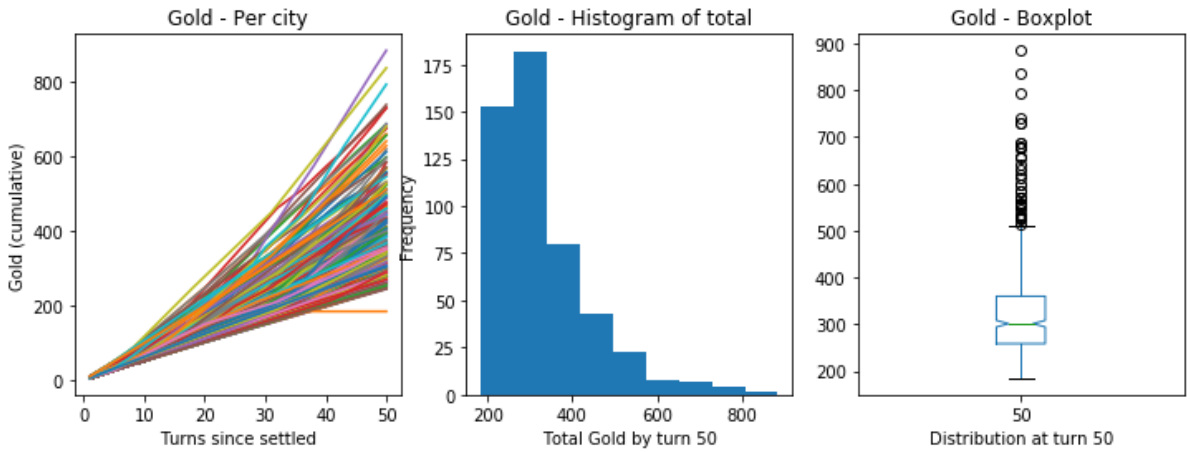
The visual representation of the Production produced looks as follows:



Gold

The cumulative Gold produced of 50 turns ranged between 184.75 and 885.00 with an average of 335.79. The distribution is right skewed like all the yield graphs.

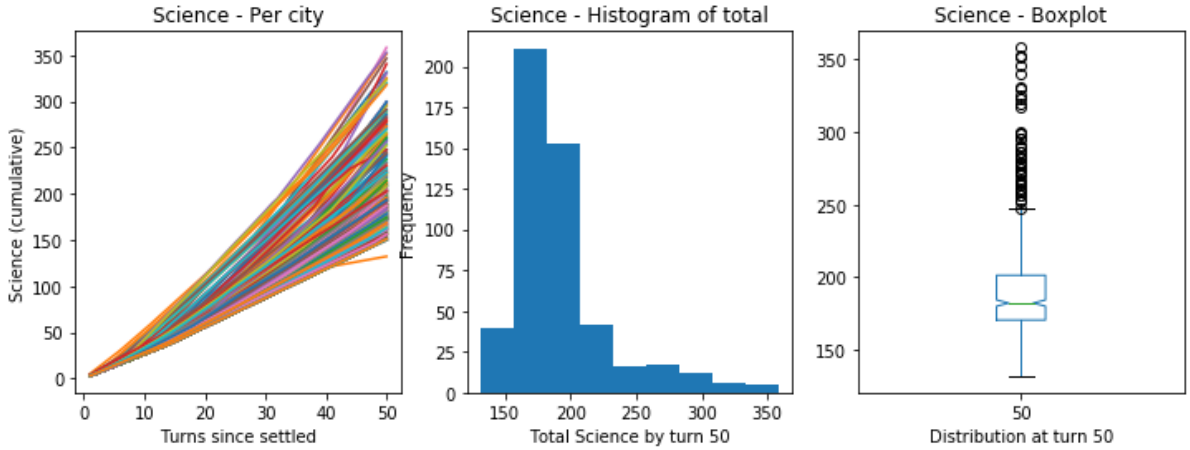
The visual representation of the Gold produced looks as follows:



Science

The cumulative Science produced over 50 turns ranges between 132.07 and 358.02 with an average of 192.41. The median is 182.28 and the histogram is right skewed.

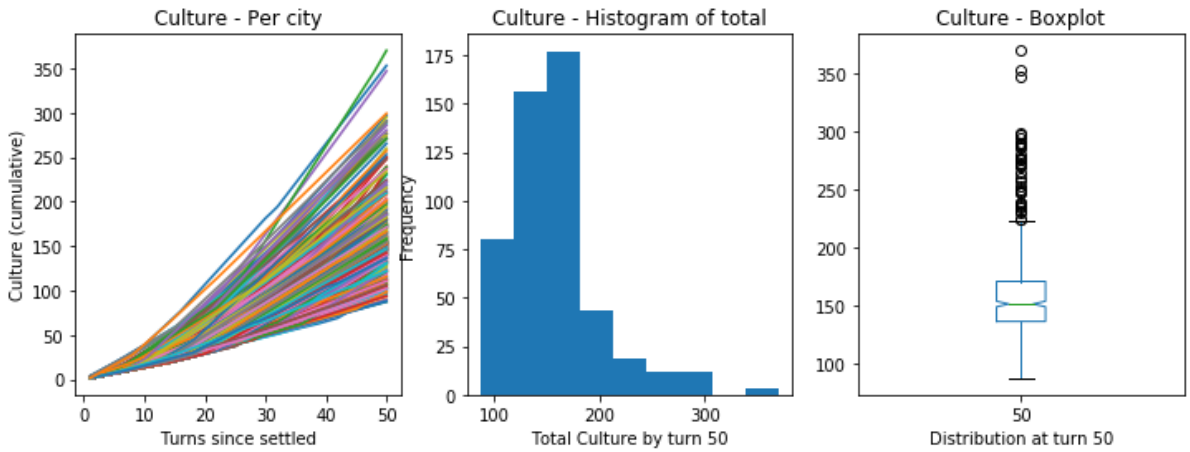
The visual representation of the Science produced looks as follows:



Culture

The cumulative Culture produced over 50 turns ranges between 87.48 and 369.91 with an average of 157.54. Like all the other yields the histogram is right skewed.

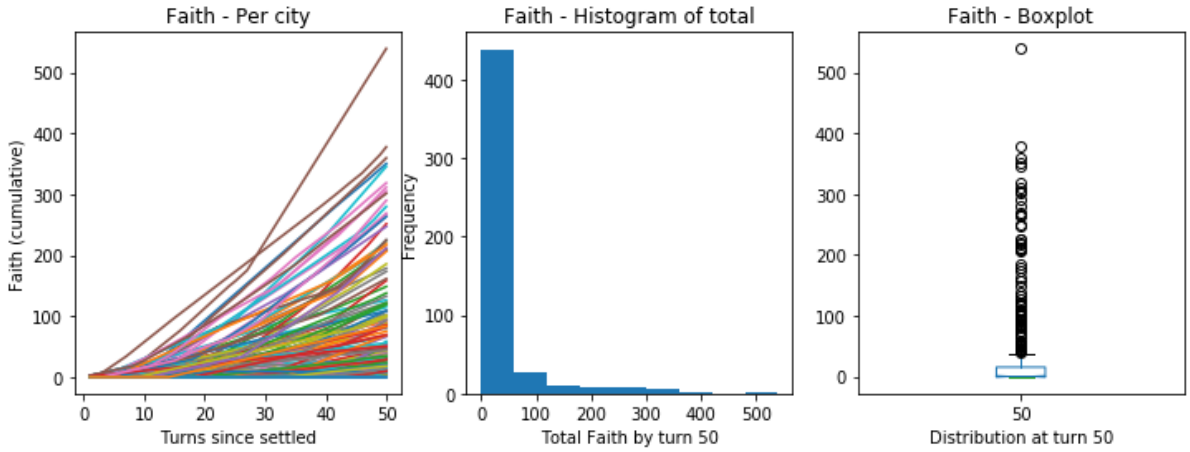
The visual representation of the Culture produced looks as follows:



Faith

Although I didn't include Faith in the cityScore metric it is still interesting to record. The cumulative Faith produced over 50 turns ranges between 0.00 and 539.01. The average is only 26.59 implying that many cities simply didn't produce much Faith.

The visual representation of the Faith produced looks as follows:



Looking at the Boxplot is informative and shows why Faith shouldn't be included in the cityScore calculation.

Other

In addition to these yields I also captured the following metrics which are included here for interest only.

Population at the end of 50 turns ranged between 2 and 8 with the average being 4.33.

Population	# of cities reached Pop	Average turn reached Pop
1	502 (100.0%)	1.0
2	502 (100.0%)	8.5
3	482 (96.0%)	19.9
4	397 (79.1%)	30.7
5	234 (46.6%)	37.7
6	79 (15.7%)	41.5
7	19 (0.04%	45.5
8	1 (0.00%)	46.0

Housing ranged between 2 and 12 with an average of 7.27.

Amenities ranged between 0 and 5 with an average of 1.51.

Amenities Needed ranged between 0 and 3 with an average of 1.41.

Classification – Prediction of good cities

After this initial analysis I used a XGBoost Binary classifier to predict good cities. I pre-determined this route, specifically as I wanted to experiment with XGBoost. However, it turns out the predictions are relatively accurate.

The confusion matrix:

Predicted	0	1	All
Actual			
0	63	13	76
1	10	15	25
All	73	28	101

The model metrics:

Accuracy	77.2%
Error	22.8%
Precision	53.6%
AUC	75.3%
Recall	60.0%
Misclassification	22.8%

I intend redoing this exercise with more data to see what impact it will have on both the model, and the explanation.

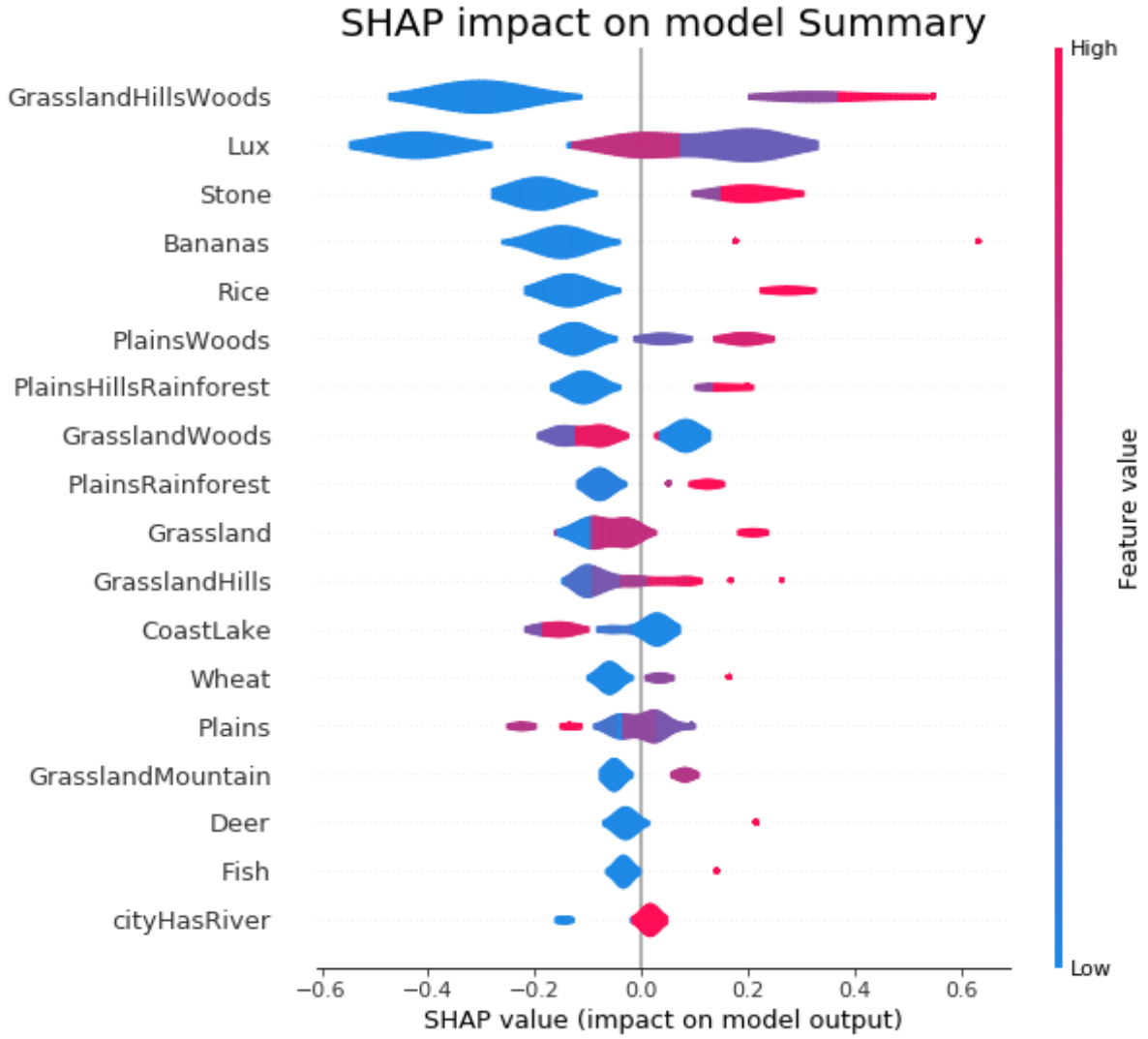
The final model was obtained by performing a randomised search cross validation against the training dataset. This dataset contained 401 cities and represented 85% of all the cities.

I used 10-fold cross-validation nested in the randomized search validation with 10 folds and 20 iterations. In effect creating and testing 2000 models per cycle.

The model was chosen using a combination of Recall and Misclassification. That is, I looked for the best recall vs misclassification trade-off.

The confusion matrix and model metrics indicate a reasonably accurate model and we can safely assume the explanation/output of the model is valid.

Model interpretation



The python SHAP module is based on Shapley values and forms the basis of an objective model explanation. While I could review the decision trees to extract the model decisions it would be time consuming and error prone.

Shapley values are concerned with how you attribute the gain from a co-operative venture proportionally and fairly among all the participants of the venture. In the scientific papers the author of SHAP argues that this is an accurate representation of the model behaviour. (You can find the papers in the Reference Documents folder of the solution.)

Interpreting the graph above the things to look for when settling a Civilization 6 Rise and Fall starting city are:

- 2 or more Grassland (Hills) with Woods tiles are great. (1 is better than none, though)
- 1x Luxury is great. (More than 1 isn't significant, not having one is significantly negative).
- 2x Stone is good. (1 is better than none though)
- Bonus resources are good and are significant in this order of preference:
    - 1x Bananas tile.
    - 1x Rice tile, more than 1 isn't significant.
    - 2 or more Wheat tiles.
    - 1x Deer tile
    - 1x Fish tile
- 2x Plains with Woods is good.
- 2x Plains (Hills) with Rainforest is good.
- Minimal, or no, Grassland with Woods tiles are preferable.
- Plains with Rainforest are positive.
- 8 or more Grassland tiles are positive. (Less than this is generally negative)
- 4 or more Grassland (Hills) tiles are positive.
- 1x Coast and Lake tile is marginally positive, more than this is negative.
- Minimal Plains tiles are preferable.
- 1 or more Grassland Mountain tiles are positive
- Not settling next to a River is negative.

Conclusion

This analysis shows that the prosperous cities can be estimated from the city centre and 18, within 2 tile distance from the city centre, tiles. The model also gives a clear indication of desirable features.

Additionally, some opportunities for feature engineering exists, especially if more data can be utilised. That is, features that are currently grouped can be separated to have a more granular model.