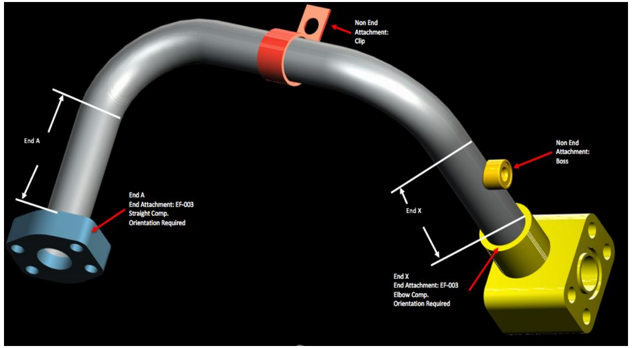
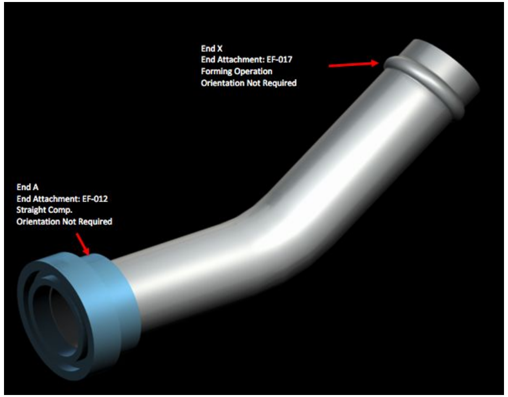
Predicting Caterpillar Tube Assembly Pricing

Karthik Venkataraman

What are Tube Assemblies (TA)?

- Caterpillar manufactures construction and mining equipment
- These equipment use complex sets of TAs for their pneumatic operations
- Each TA is made of one or more components
- TAs can vary in base materials, number of bends, bend radius, bolt patterns, and end types
- Caterpillar relies on a variety of suppliers to manufacture these tube assemblies, each having their own unique pricing model





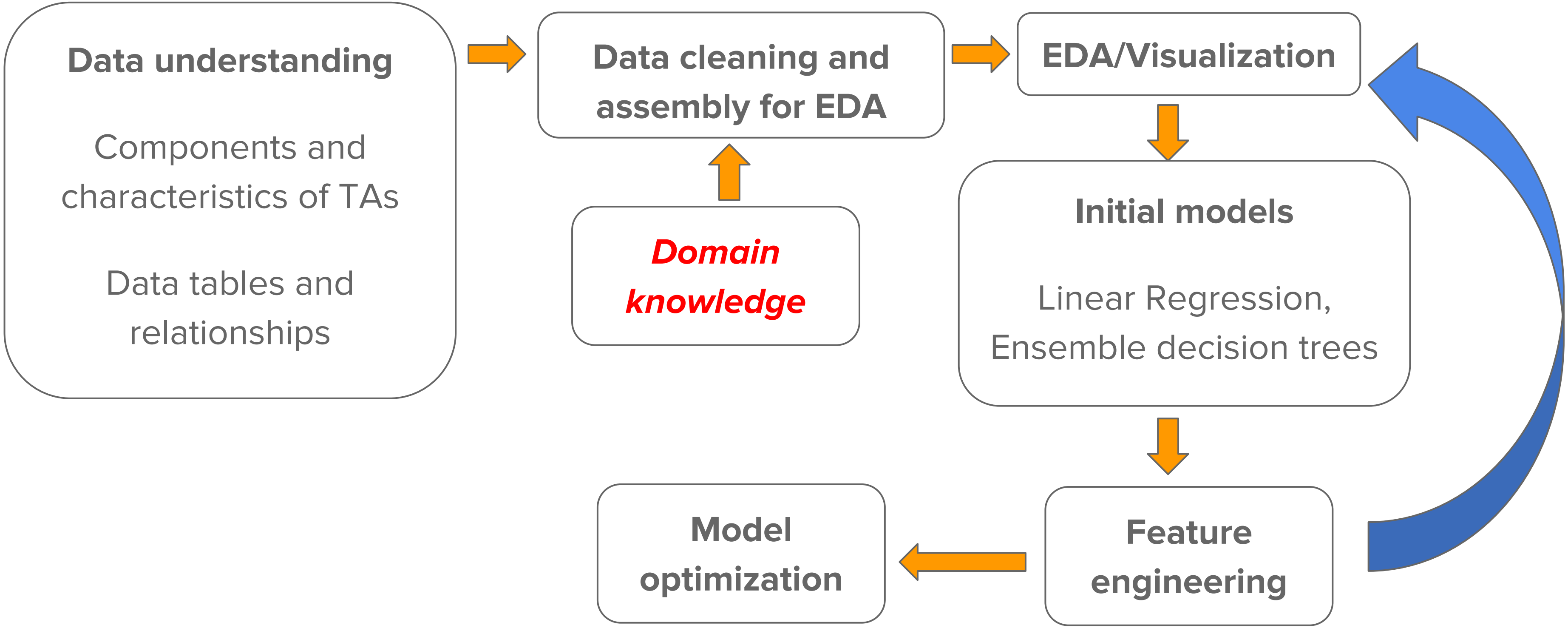
Problem statement

1. Given detailed tube, component, and annual volume datasets, develop a model to predict the price that a supplier will quote for a given tube assembly … ***goal of Kaggle competition***

And more importantly,

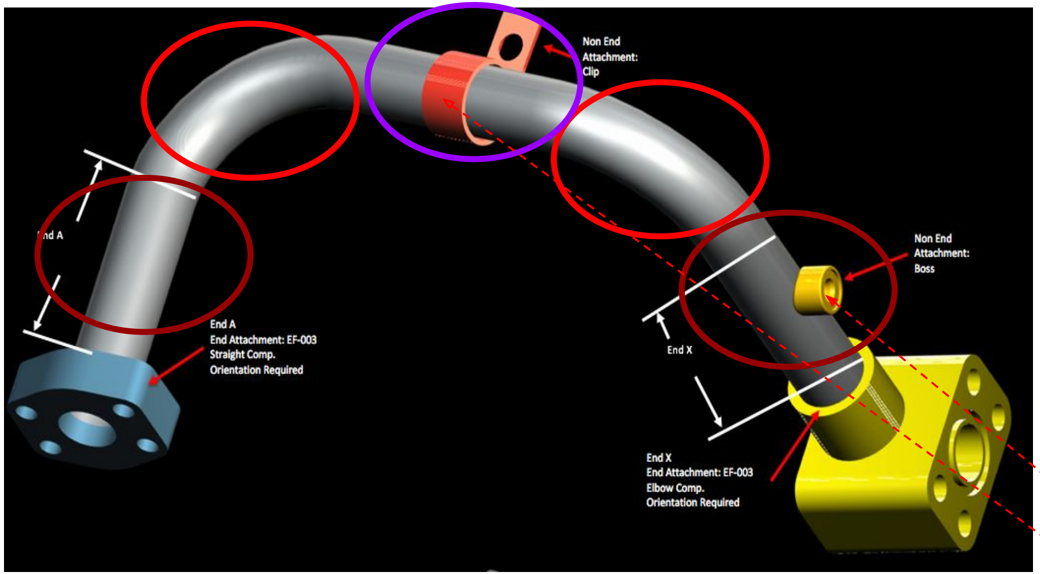
2. Determine the top features that contribute to TA pricing

Study methodology



Data understanding. Hypotheses on key features.

Understanding TA components: Example



- 5 components
  - 2 90-degree elbows, with tig welded ends
  - 1 straight connector, with tig welded ends
  - 2 straight connectors, 1 end tig welded, 1 end threaded
- 4 “other” components
  - 1 end X attachment
  - 1 end A attachment
  - 1 boss
  - 1 bracket

Data tables and key columns

TA pricing
TA ID
Annual usage
Min. order qty
Bracket pricing
Qty
Cost

TA Bill of Materials
TA ID
Component 1-8 ID
Qty each component

Component type
Component ID
Component descr.
Component Type ID

TA characteristics
TA ID
Material ID
Dia, Wall, Length
boss, # bracket
End fitting “A” ID
End fitting “X” ID
End fitting L <1-2x dia

TAs

- 30,213 data points total
- 8,885 unique TAs
- 2,048 unique components
- 29 unique component types (like elbows, bolts, flanges etc)
- 27 unique end fittings (for ends “A” and “X”)

Pricing

- Bracket pricing, based on quantity
- Non-bracket pricing, based on minimum order quantity and annual usage
- 57 suppliers, with most supplying unique TAs (i.e. **no competition**)

Other data tables

End fitting
End form ID
End forming (Y/N)

Component type
Component type ID
Description

Component connection type (ex. Flare angle, thread pitch)

Component connection end form (ex. Threaded male, braze welded)

Component characteristics (separate table for each type)
Component ID
Component Type ID
Unique component characteristics (ex. Orientation, bolt pattern, connection type, connection end form)

Component specs
Component ID
Spec ID

Thoughts on approach

- ❏ There are potentially tons of features available to build a model
- ❏ Using every possible feature would:
  - ❏ Require extensive manipulation, cleaning and joining of tables
  - ❏ Potentially lead to increased model complexity, overfitting and poor predictability to new data sets
- ❏ ***Identifying key potential features using previous knowledge (and verifying using EDA) is a must to keep project scope in check. Even with selected features, techniques like regularization and PCA are probably needed to prevent the model from overfitting***

Typical components of (supplier quoted) price

Quoted Price ~ Fixed cost (FC) + Variable cost (VC) + Margin (M)

<u>Note</u>
Quoted price is denoted as "Cost" in the dataset

Supplier Variable cost

- ❑ Covers everything that scales directly with number of units produced
- ❑ Ex: steel and other materials, direct labor, utilities (air, electricity, water) etc

Supplier Fixed cost

- ❑ Covers all costs other than variable cost
- ❑ Ex: R&D support, new equipment installation, equipment setup time, sales and general administration, factory/office maintenance etc

Margin

There are a number of factors that could affect how a supplier sets margin for a quoted price. For example,

- ❑ Number of suppliers bidding for a particular TA
- ❑ Length of time Caterpillar has a relationship with supplier at time of bidding. Supplier could choose to set a higher margin if they know that switching costs for Caterpillar are higher because qualifying a new supplier takes time, or if they have specialized capabilities
- ❑ Annual expected sales volume (across all TAs) to Caterpillar
- ❑ How they prioritize topline (revenue) and bottomline (profit) growth at time of quote

Hypotheses regarding key features

- (VC, FC) Factory fixed costs will be spread across more units. Price (per TA) should be inversely proportional to:
  - TA quantity ordered
  - Minimum order quantity
  - Annual usage
- (VC) Tube (or component) material of construction. Ex. 316L > 304 stainless steel price
- (VC) Number and **type** of each component in a TA. Ex. Tig welded > braze welded price
- (VC, FC) Number of specifications/tolerances that each component needs to meet. Ex. Price of components with dimensional tolerance specs > no specified tolerances
- (VC, FC) TA end lengths that are < 1-2x tube dia require special tooling = higher price

Other questions to be explored during EDA

- ❏ Separate models for bracket and non-bracket pricing?
- ❏ (M) Price = f(Supplier)?
- ❏ (VC) Price = f(Quote date)? Ex. steel commodity pricing dynamics affecting all TAs
- ❏ Do the unique characteristics of each component need to be included as features in a model? Or are higher level features (ex. the 29 unique component types) sufficient?
- ❏ (M) Price = f(Annual usage across all TAs)?
- ❏ (M) Price = f(Length of supplier relationship with Caterpillar)?

Data cleaning, assembly and EDA, #1

Steps

Note: Not all features discussed in the hypotheses were included in this round of analysis

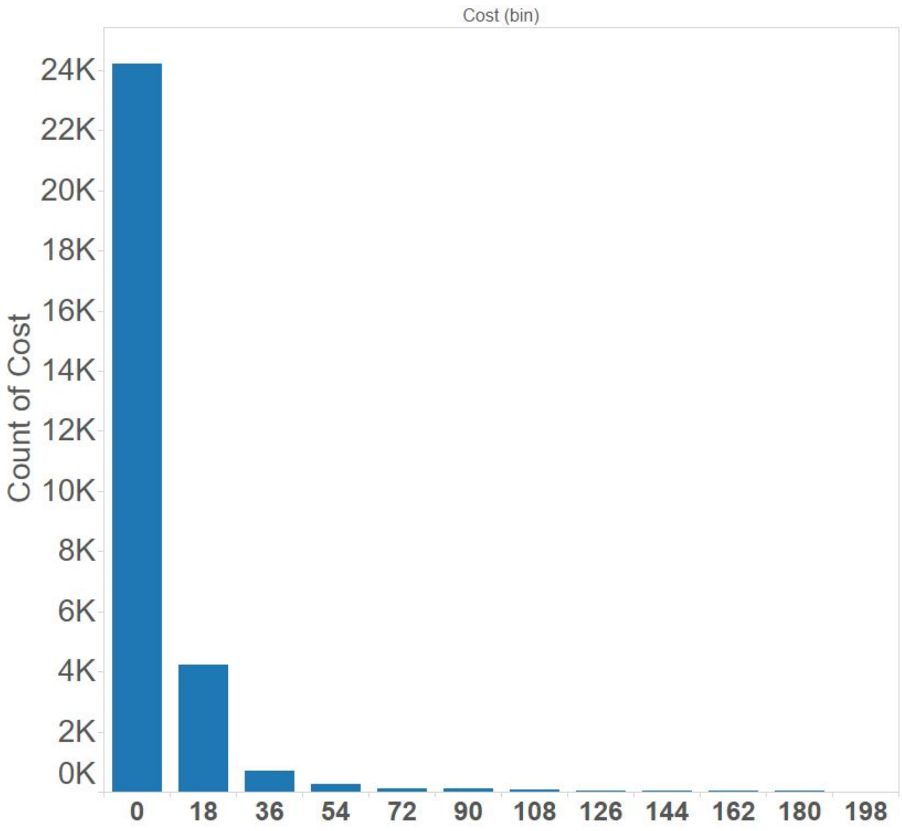
1. Merge TA pricing and TA characteristics datasets
2. Calculate number of components per TA from bill of materials dataset, and merge with (1)
3. Explore in Tableau

TA pricing
TA ID
Annual usage
Min. order qty
Bracket pricing
Qty
Cost

TA characteristics
TA ID
Material ID
Dia, Wall, Length
boss, # bracket
End fitting "A" ID
End fitting "X" ID
End fitting L <1-2x dia

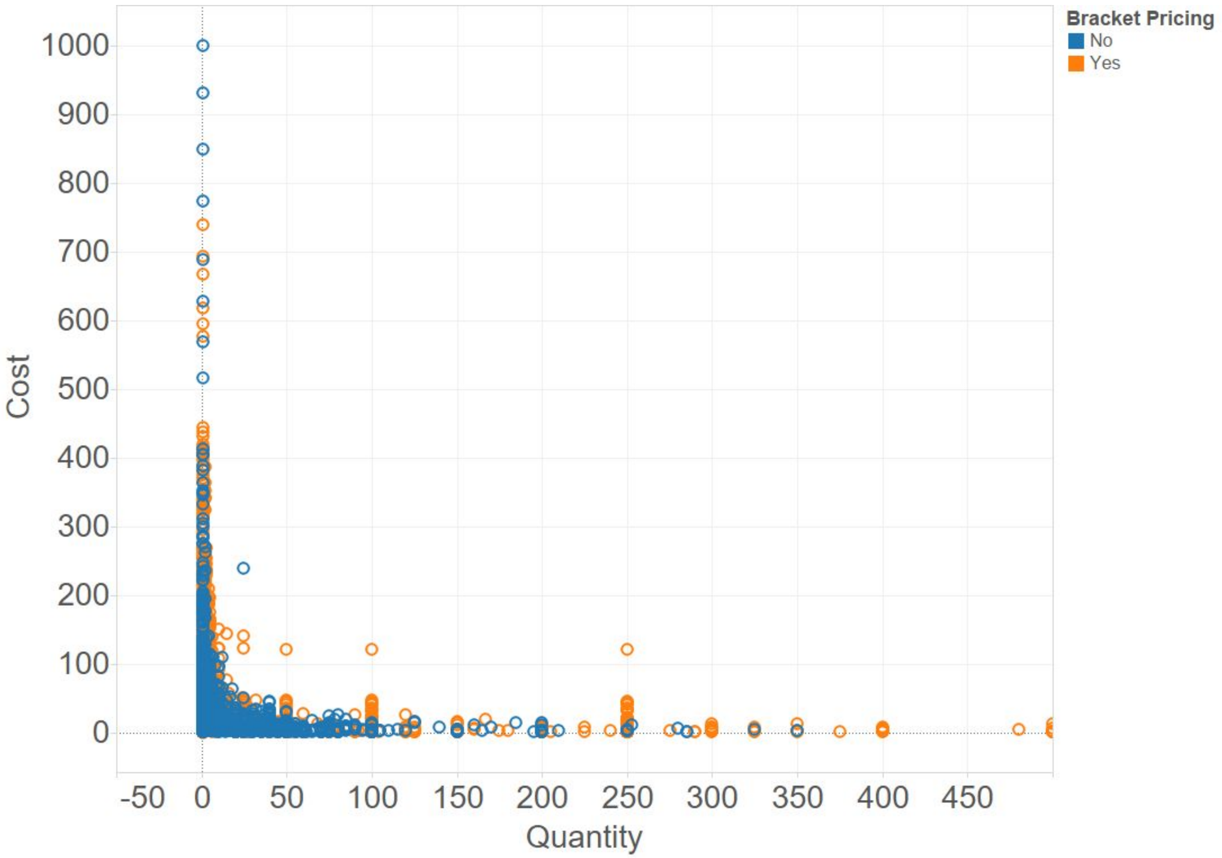
TA Bill of Materials
TA ID
Component 1-8 ID
Qty each component

Cost is positively skewed … most costs are <$20



Note
Cost bins > $198 not shown

Quantity strongly affects pricing (bracket and non-bracket)



Lower and upper bounds of model performance

Metric: Root Mean Squared Log Error (RMSLE)

- ❑ Differences are expressed as LN(pi+1) - LN(ai+1), instead of (p-a) in RMSE, where pi is predicted value, ai is actual value, and LN is the natural log

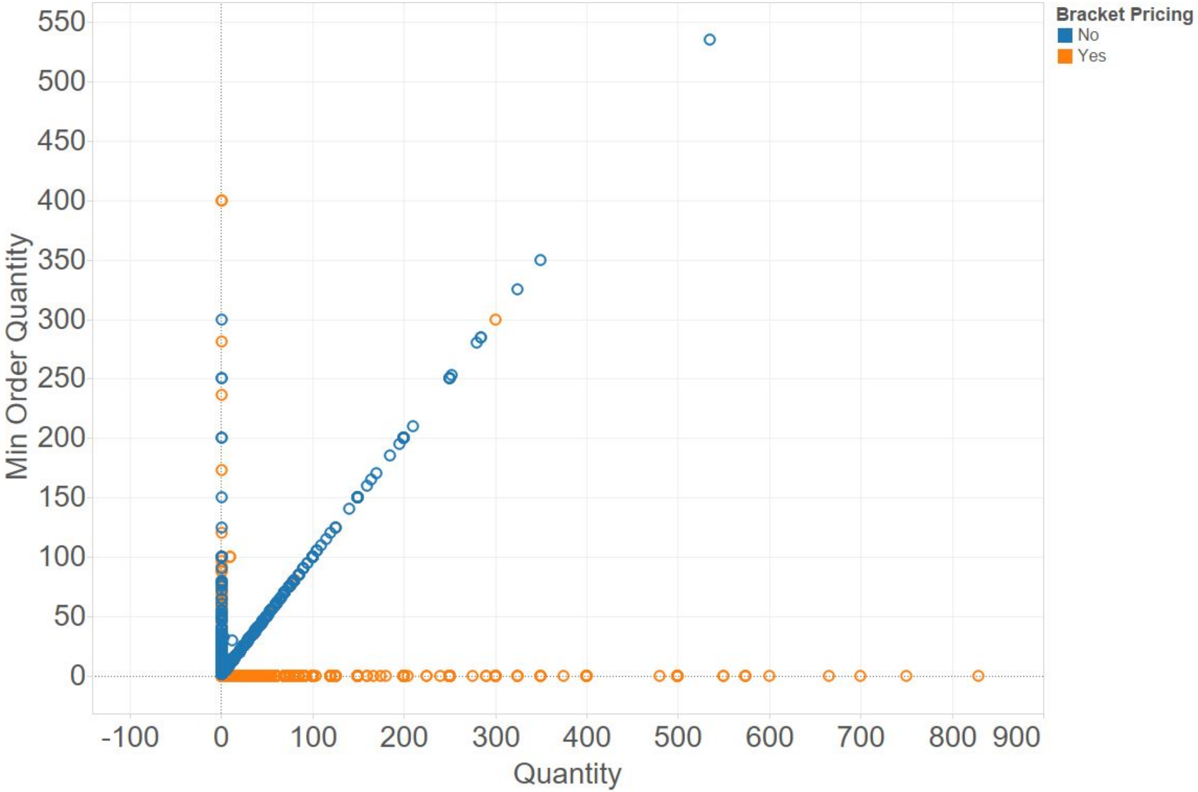
Lower bound

- ❑ If the cost was modeled as the mean cost of the entire dataset, then **RMSLE = ~0.95**

Upper bound

- ❑ Need to estimate using data from same TA, either from the same supplier who has provided quotes over time, or from multiple suppliers
- ❑ From a limited dataset of 2 TAs, each of which was quoted 4 different times by the same supplier, **RMSLE = ~0.08**

Min order qty and quantity should be the same ... but are not. This needs to be corrected before further analysis

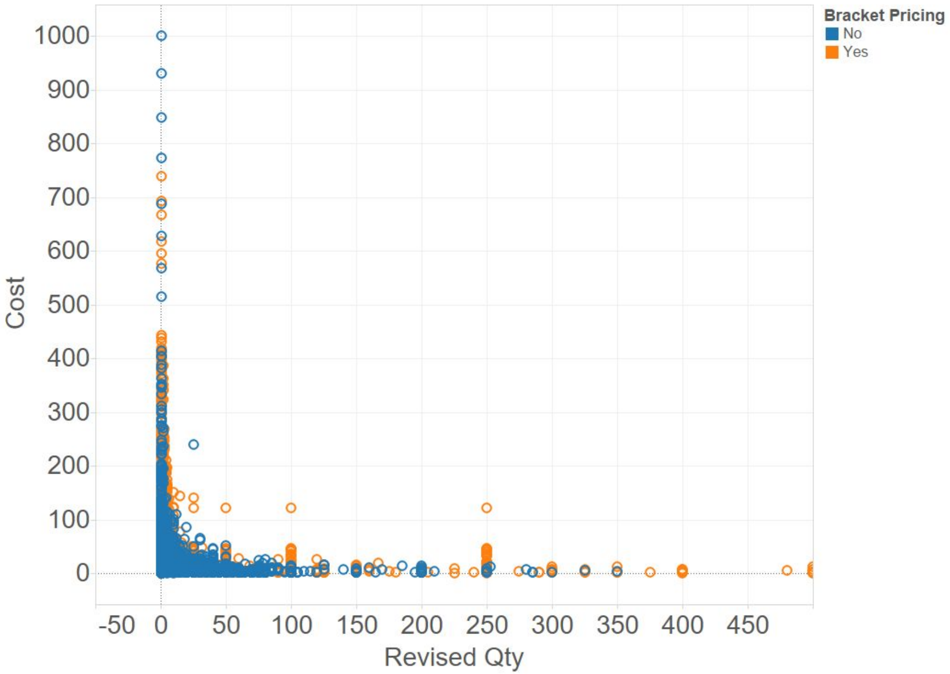


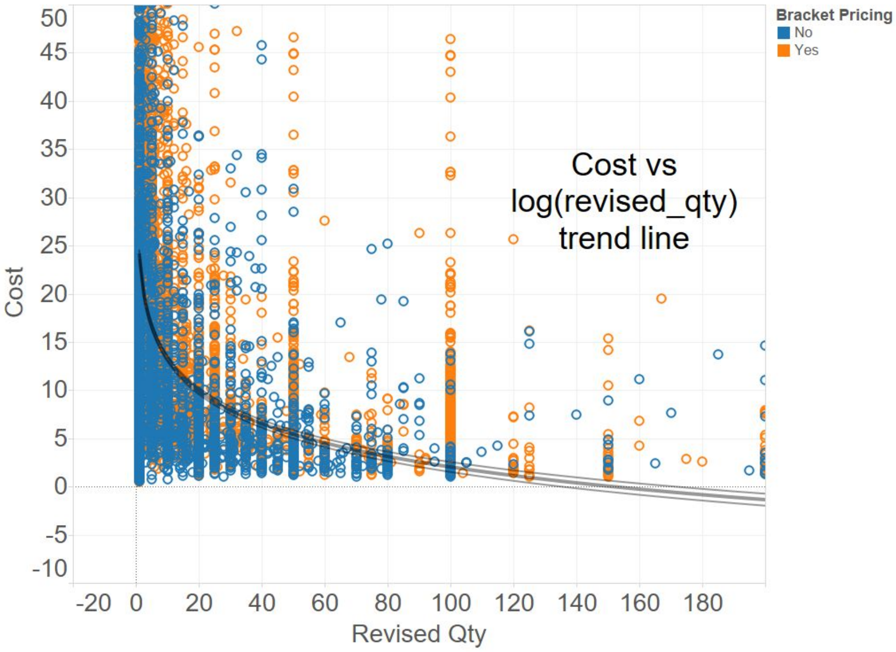
Calculating “revised qty”

- Non-bracket pricing (3,414 data points)
  - In ~76% of the cases, minimum order quantity > quantity
  - Set Revised qty = Minimum order quantity

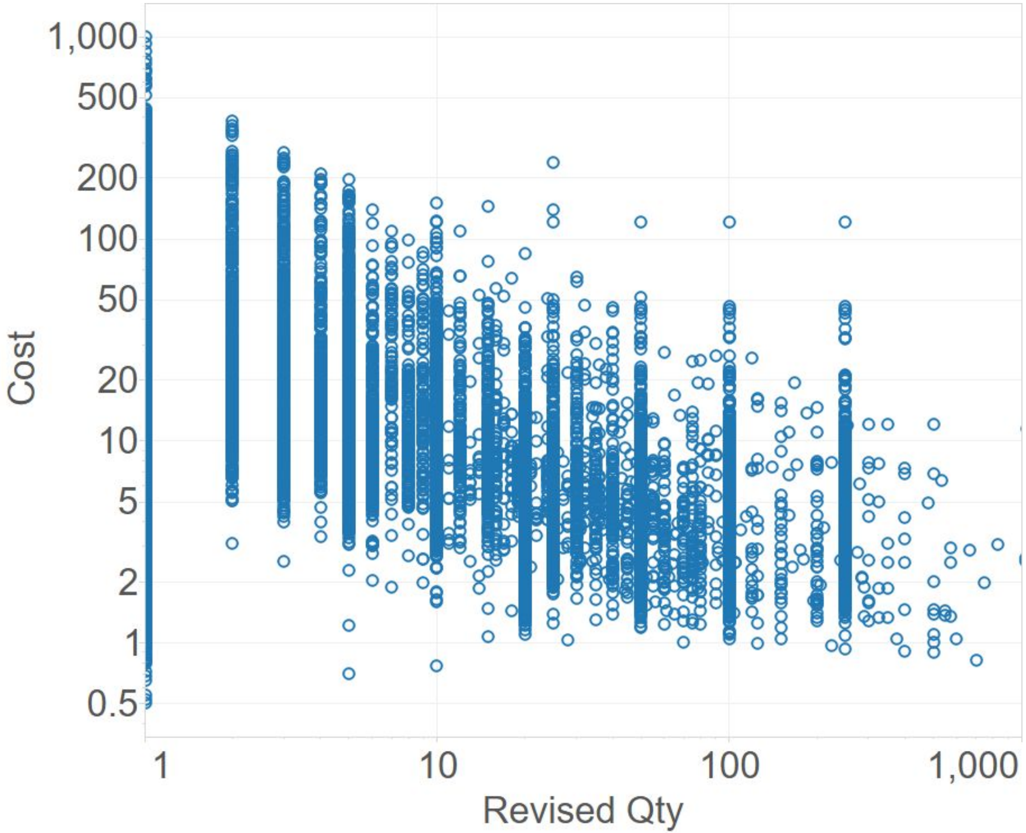
- Bracket pricing (26,799 data points)
  - In ~83% of the cases, minimum order quantity = 0
  - Revised qty = Quantity where Quantity >= Minimum order quantity
  - Delete rows where Quantity < Minimum order quantity (243 out of 30,213)

As before, quantity (now "revised quantity") affects price

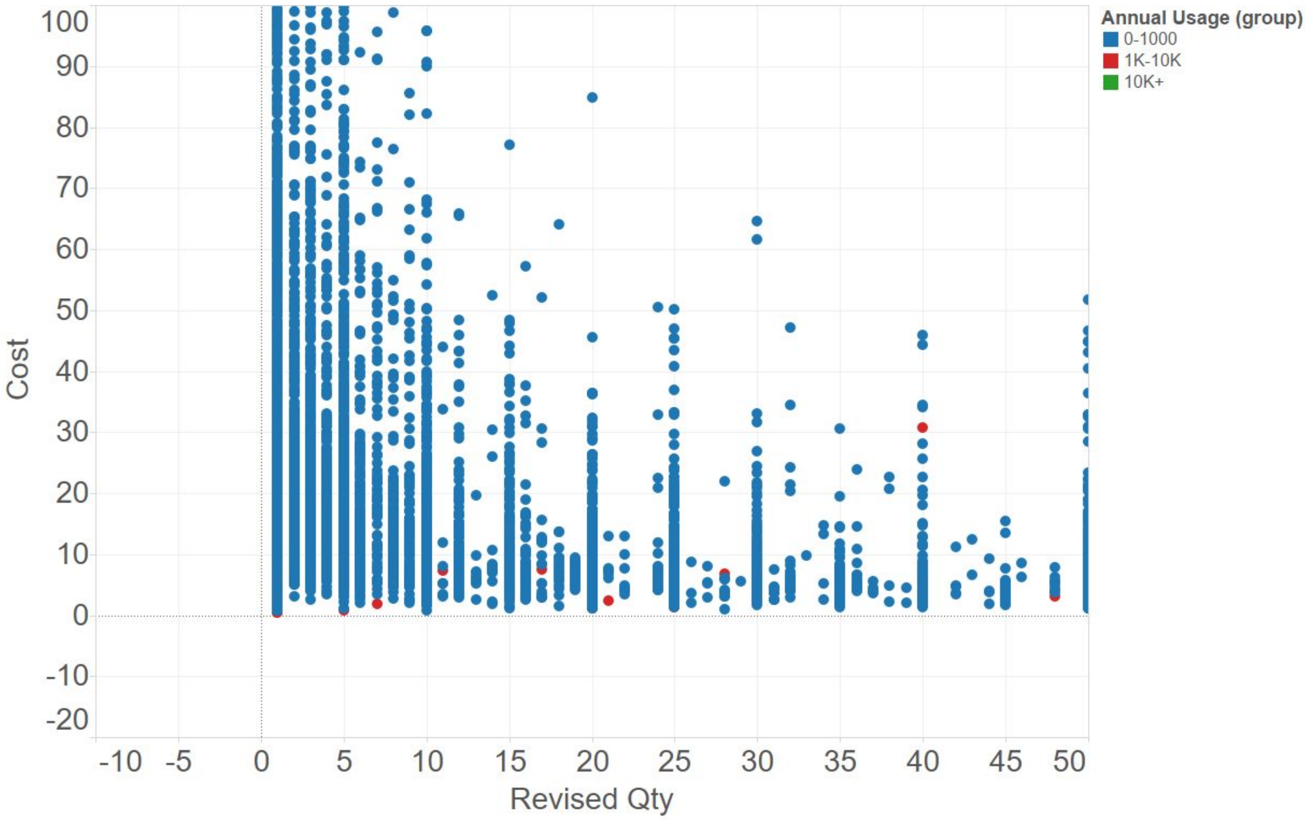




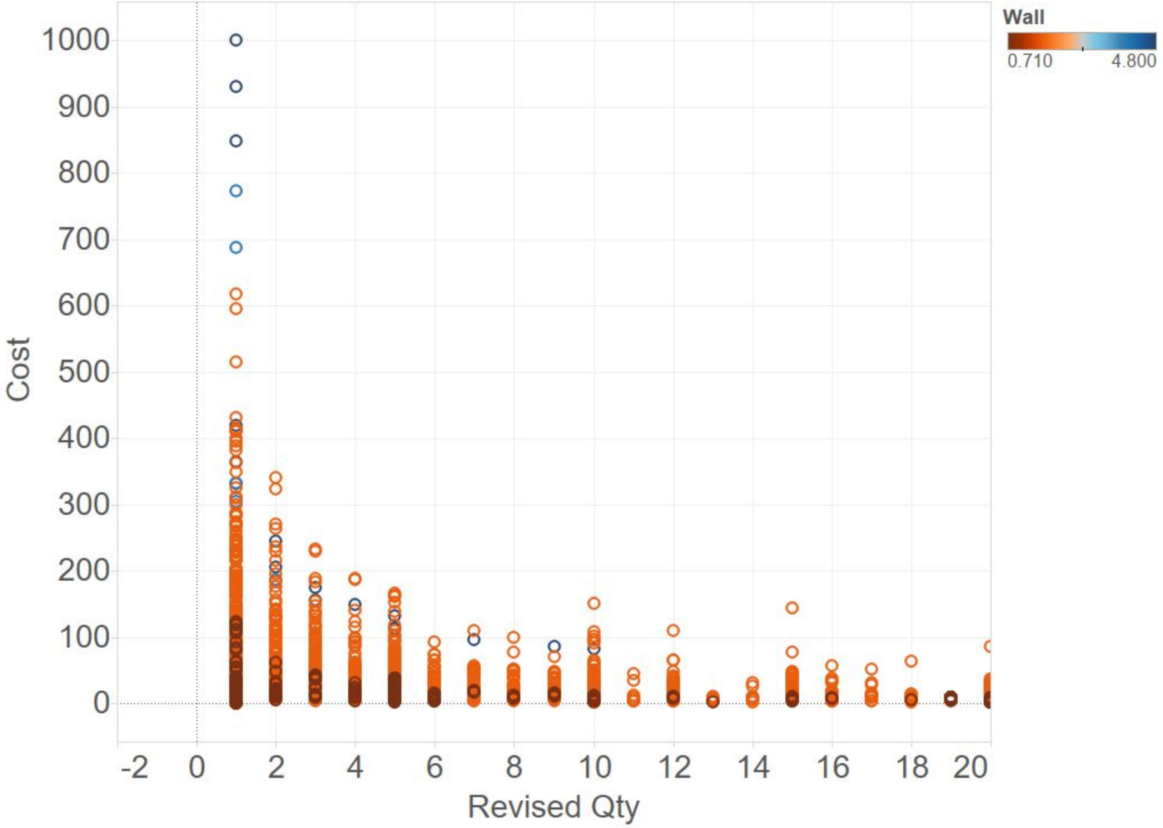
Log transformation of quantity and price may be useful for linear models



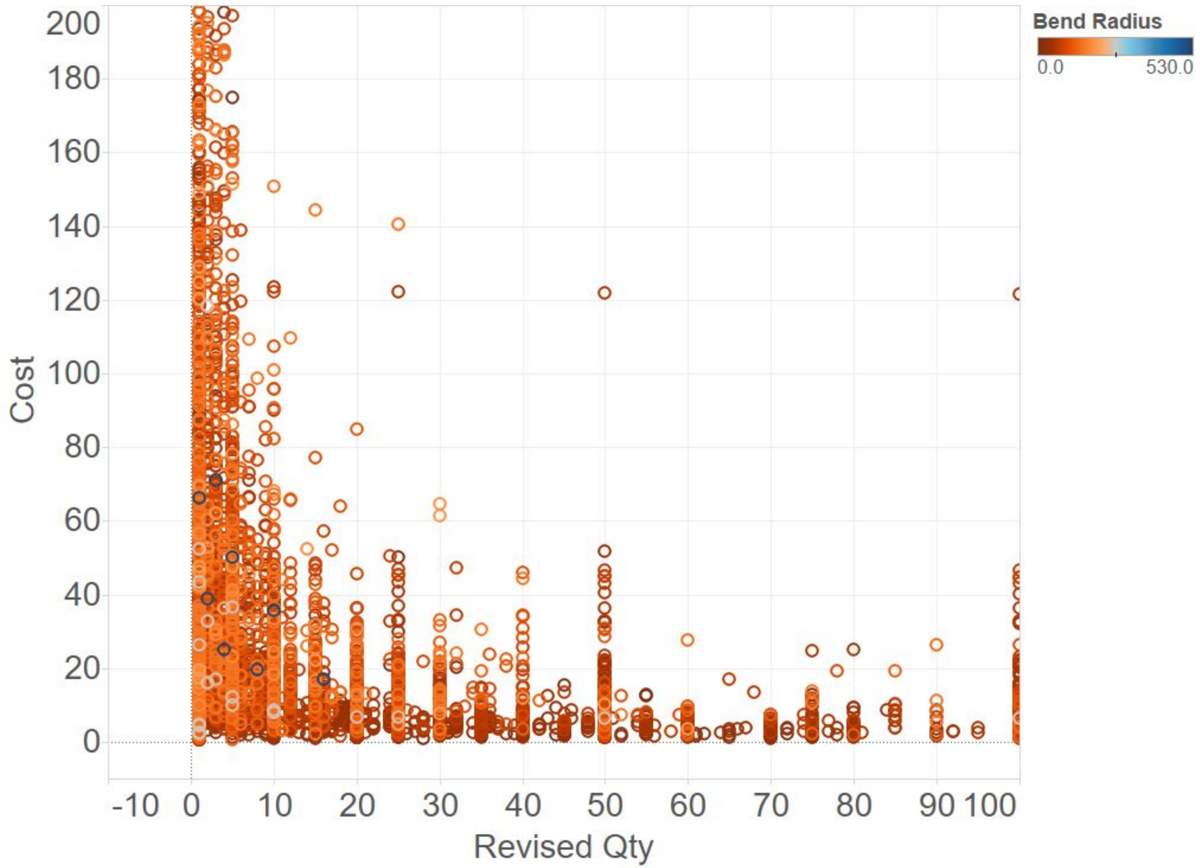
Annual volume may have an effect



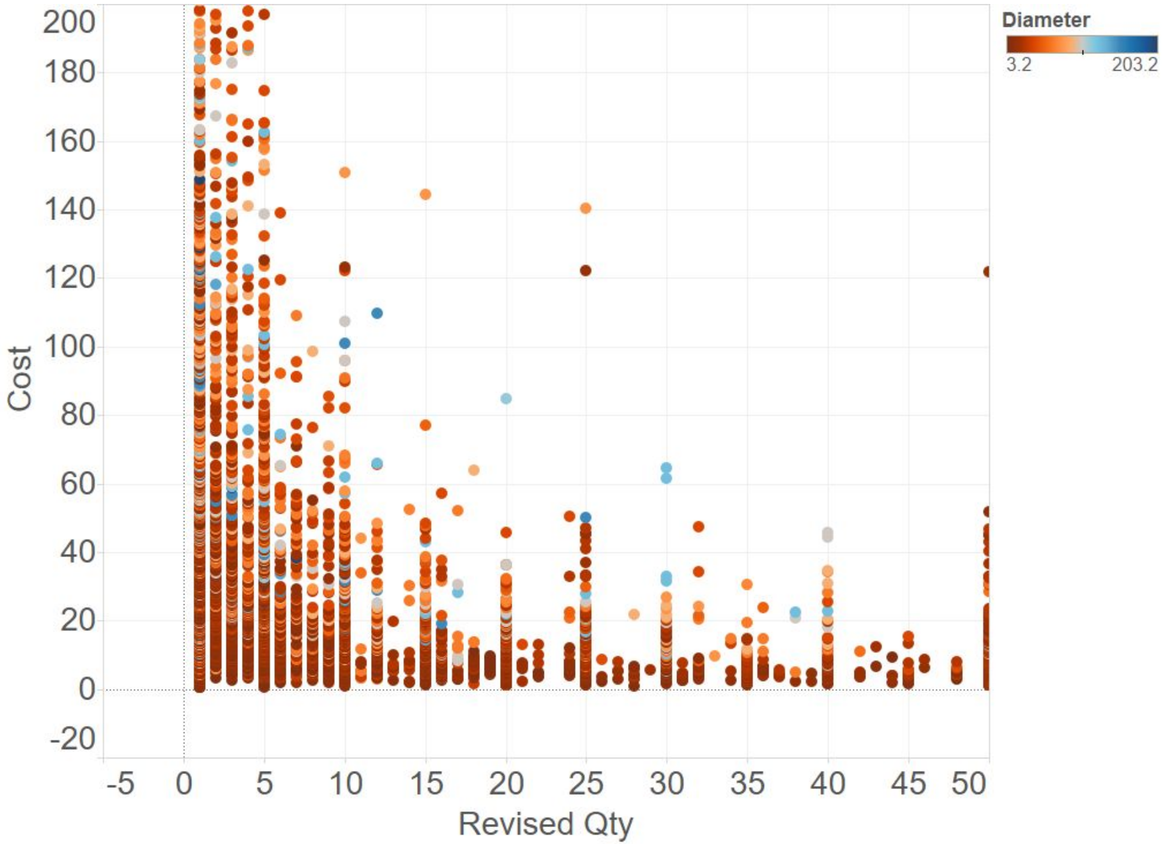
Higher wall thickness TAs seem to be priced higher



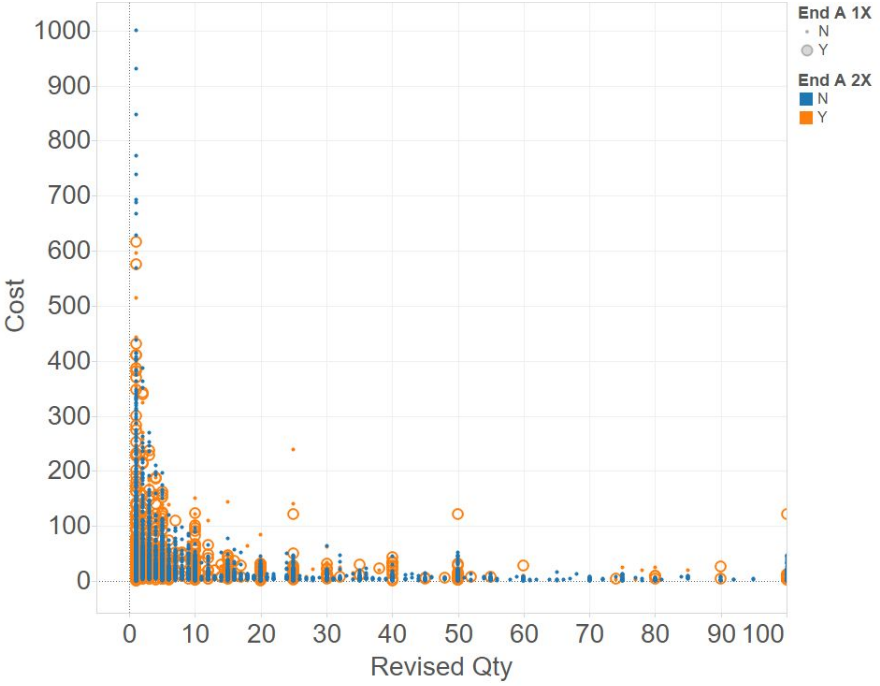
TAs with higher bend radius (i.e. “larger” parts) seem to be priced lower

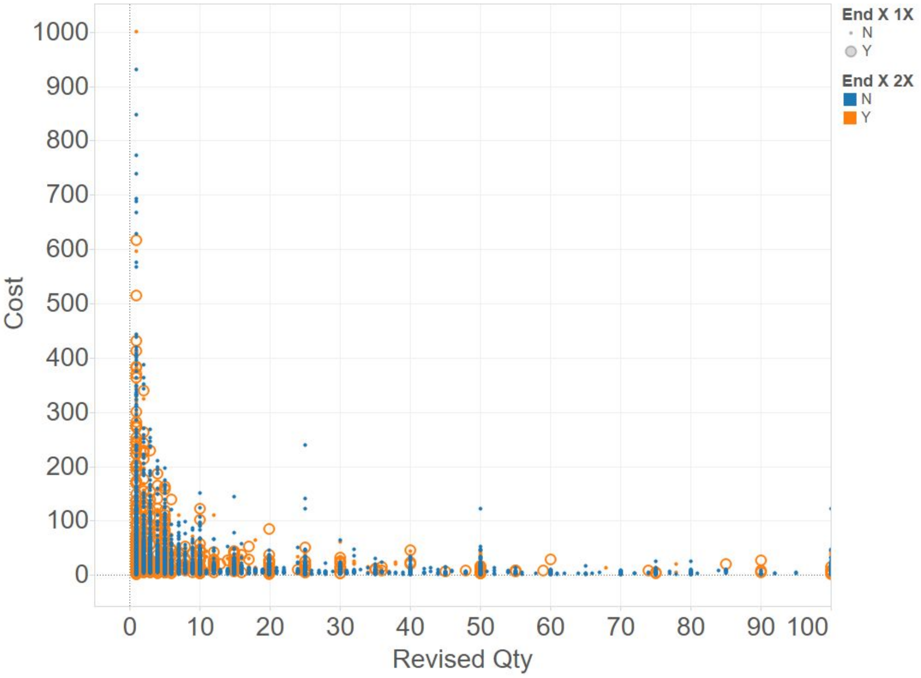


Higher diameter TAs may be priced higher, but the effect is not clear

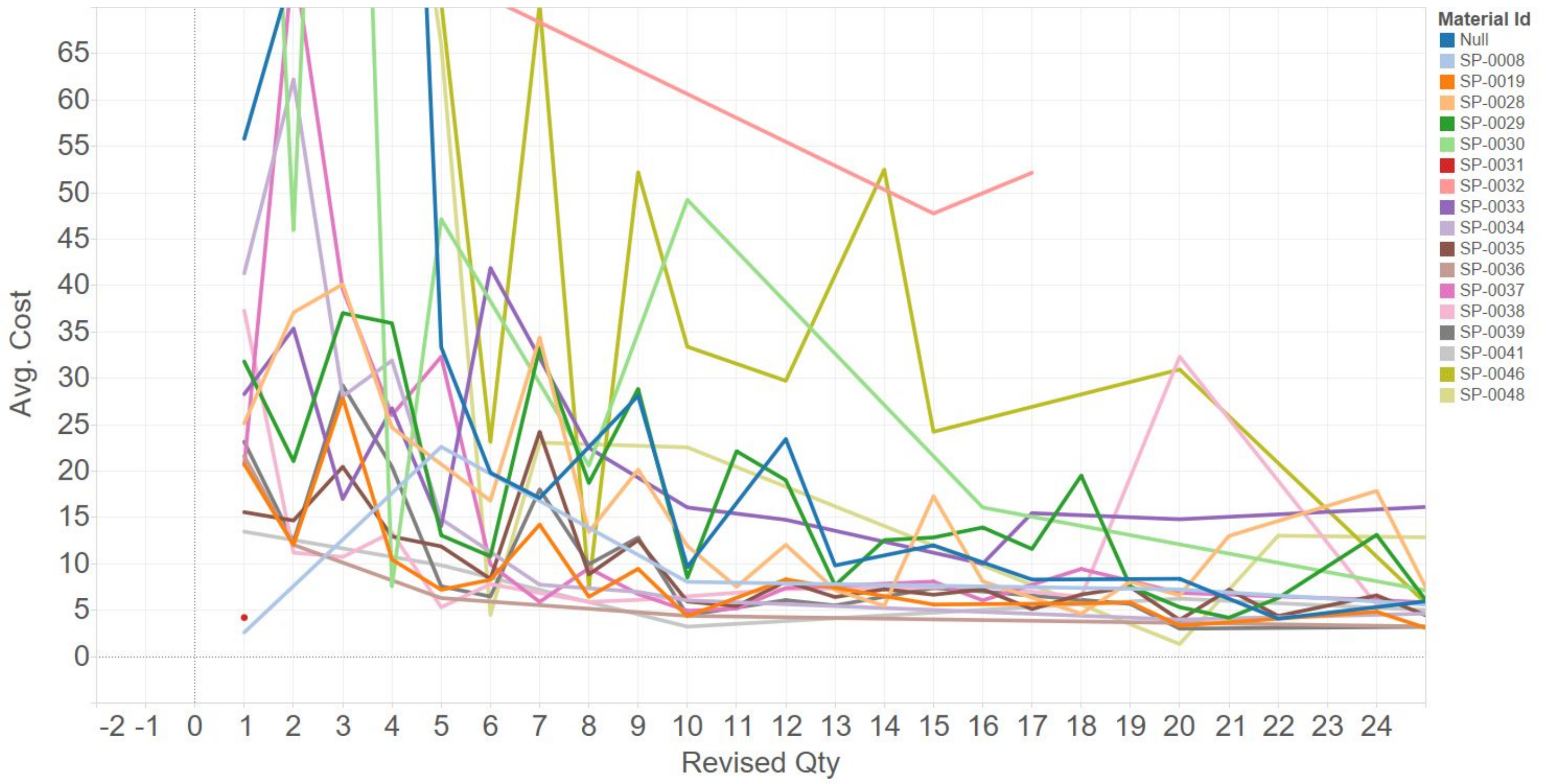


Ends “A” and “X” lengths < 1-2x TA dia doesn’t seem to affect pricing

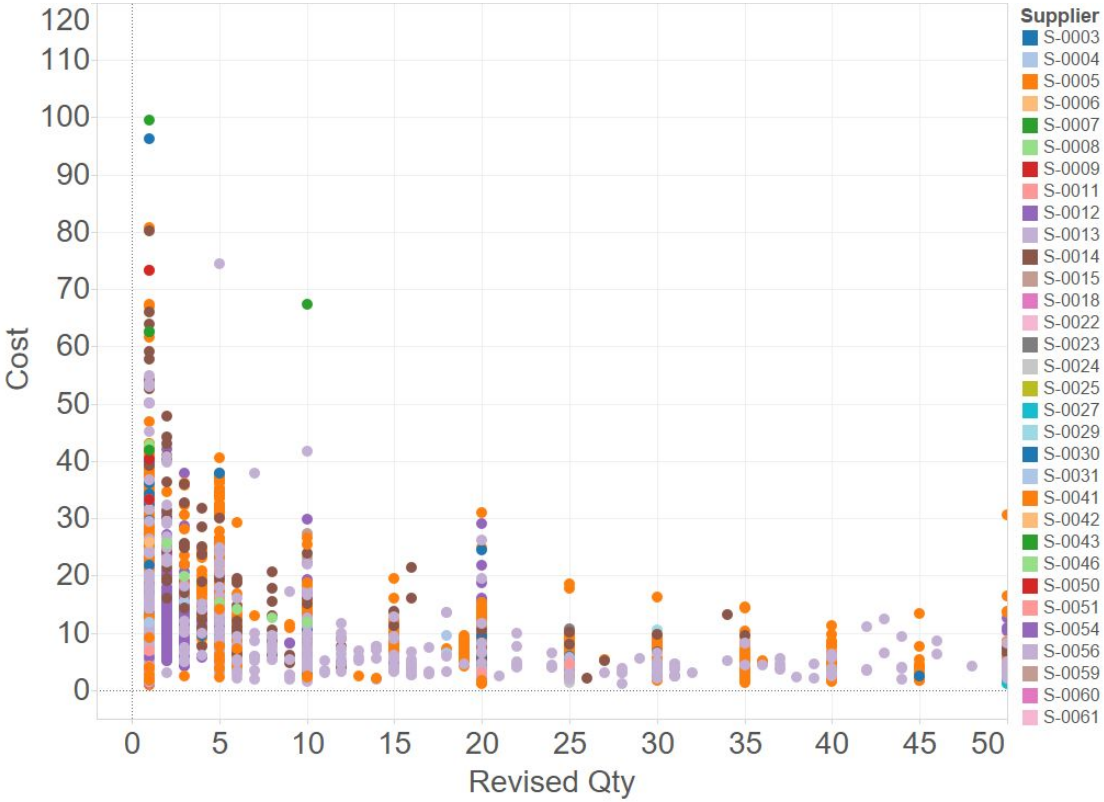




It’s unclear if material of construction affects pricing



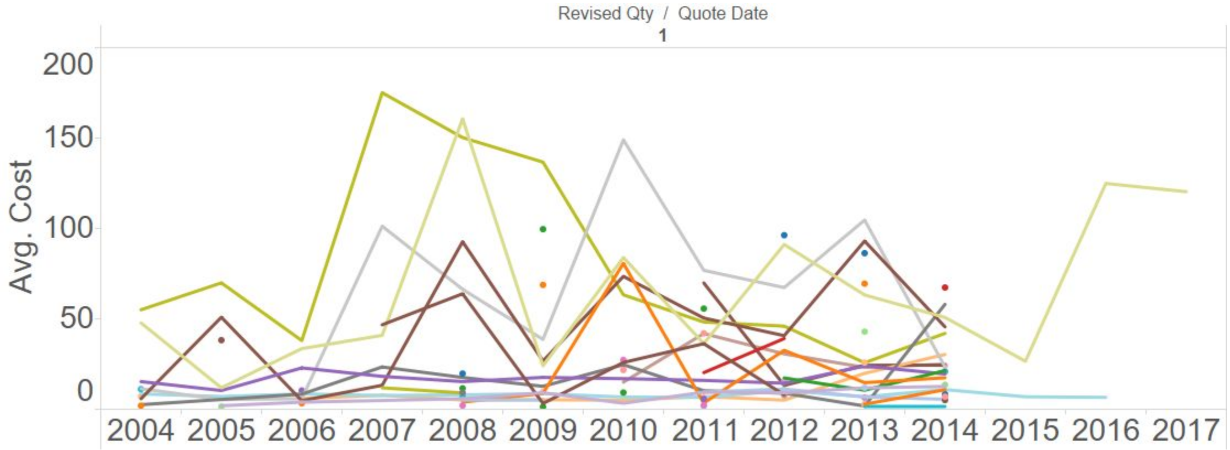
Most supplier have similar pricing dynamics (vs qty)

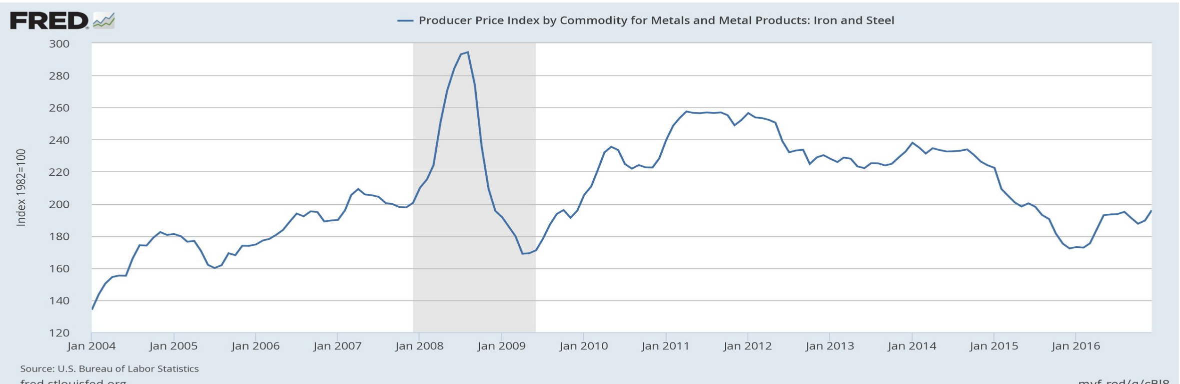


Note

Not all suppliers shown

Avg. cost (for qty = 1) for TAs made with SP-0029 material from different suppliers, from 2013-2015





- Different suppliers = different dynamics (with time)
- Don't see obvious effect of commodity prices, esp. on "low" cost TAs
- Unlikely that prices were affected by commodity prices, which may be a small component of the suppliers' overall cost

Other data cleaning, #1

Data prep for modeling

Feature/Issue	Action	Data reduction
material_id NaN	Drop NaN	0.7%
material_id categories (SP-xxxx)	Change to dummy variables	--
End lengths 1-2x (Y, N)	Change to dummy variables	--
End A and X type (EF-xxx)	Change to dummy variables	--
bend_radius = 9999	Change to 0 (all other straight components denoted as 0)	--
length = 0	Drop rows	0.09%

Modeling, #1

Models

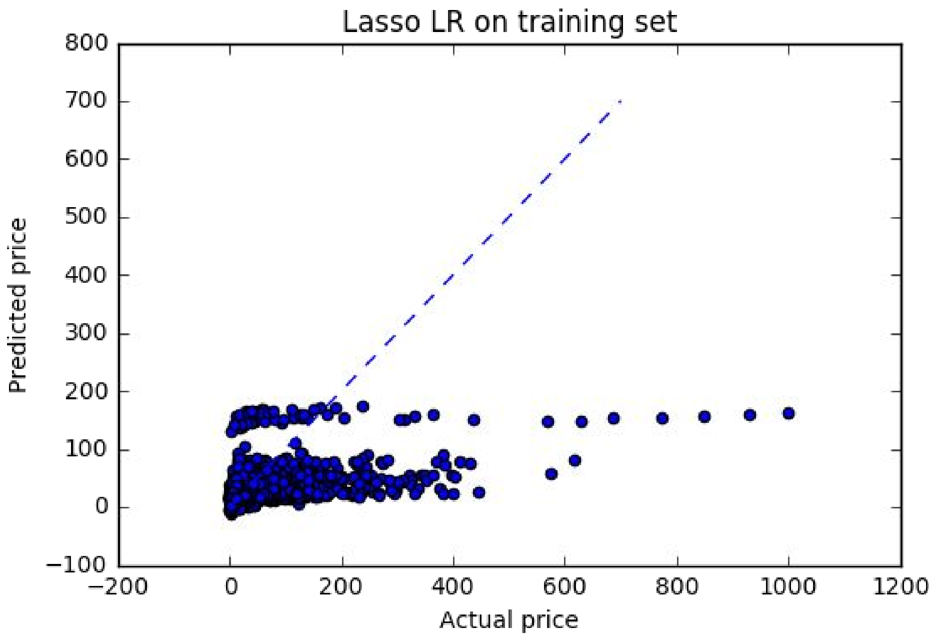
- 2 classes of models were used in this round:
  - Linear models, specifically Linear Regression
  - Ensemble tree based models, specifically Random Forest and Gradient Boosting
- The dataset was split into training (70%) and test (30%) sets
  - Cross-validation was performed on the training set
- L1 Regularization/Lasso was applied to Linear Regression to prevent overfitting … not all features likely contribute per the EDA, and we want to discard them
- No feature selection was done at this stage for the ensemble models (they are somewhat robust to correlated features and overfitting because the trees get built with different features inherently)

Input features

- All features were standardized prior to use in the models
- The following features were used
  - Revised quantity (log transformed for linear regression)
  - Annual usage
  - TA diameter, TA wall thickness, TA length, bend radius, number of bends
  - Total number of components in TA; number of boss', brackets, and other components
  - End "A" and End "X" length (<1-2x dia)
  - TA material of construction ID

Linear Regression (LASSO regularization)

	R2
Training set	0.285
CV (5-fold)	0.274 +/- 0.03
GridSearch (best alpha = 0.1)	0.284

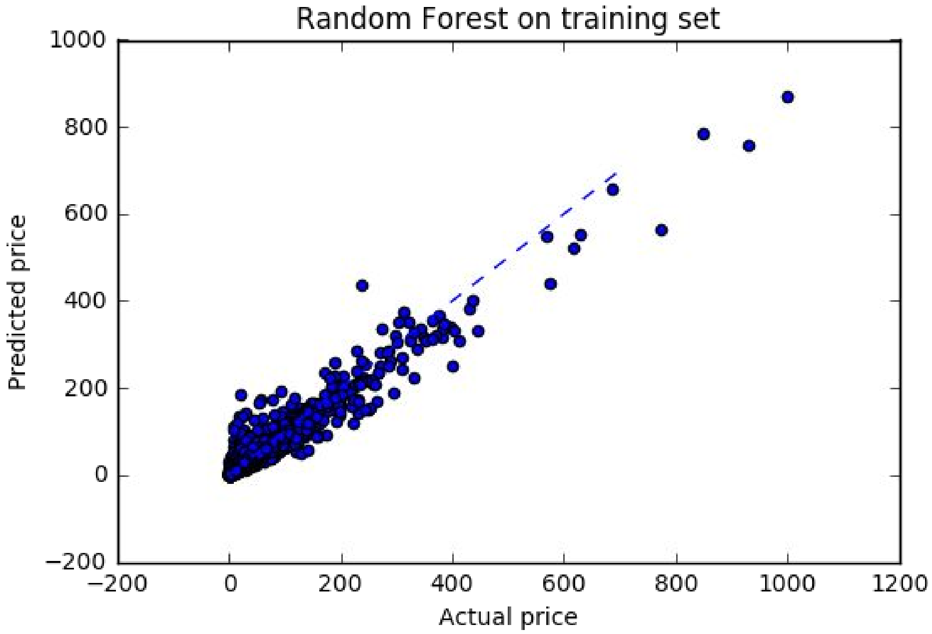


Ridge regularization gave similar results. Using log(price) as target variable also gave similar results

Linear regression doesn't perform well, even with optimization. There is a Bias problem. A linear model may not work well without a lot of additional features.

(Base) Random Forest

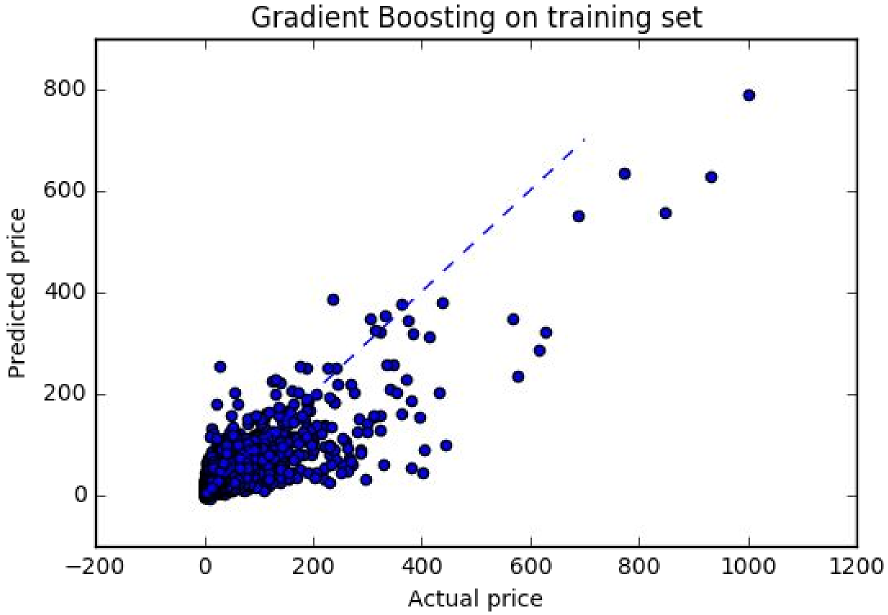
	R2
Training set	0.941
CV (5-fold)	0.631 +/- 0.057



Random Forest performs much better than Linear Regression. It suffers from a variance problem, so model tuning may be needed

(Base) Gradient Boosting

	R2
Training set	0.699
CV (5-fold)	0.577 +/- 0.078



Gradient boosting also performs much better than Linear Regression. There is potentially both a bias and variance problem, so both model tuning and additional features may be needed

Summary, #1

Key take-aways

- ❑ **Model is likely non-linear**
- ❑ Both Random Forest and Gradient Boosting perform much better than Linear Regression, and will be used for the next round of evaluations
- ❑ Linear regression will not be considered any further … given the really low scores, it’s unlikely that performance will improve with addition of features, unless the “right” features/transformation of features are discovered
- ❑ Features that provide additional discrimination of prices at the lower end may help

Data cleaning, assembly and EDA, #2

Additional features

Let’s explore if additional features help. The following features (from our hypotheses) were considered in this round of analysis:

- ❑ Number of specifications for every TA
- ❑ The number of every type of component in the TA (ex. 2 elbows, 4 sleeves etc)
- ❑ Whether the TA ends are formed or not

Data table prep

FROM ROUND 1

TA characteristics
TA ID
Material ID
Dia, Wall, Length
boss, # bracket
End fitting “A” ID
End fitting “X” ID
End fitting L <1-2x dia

TA pricing
TA ID
Annual usage
Min. order qty
Bracket pricing
Qty
Cost

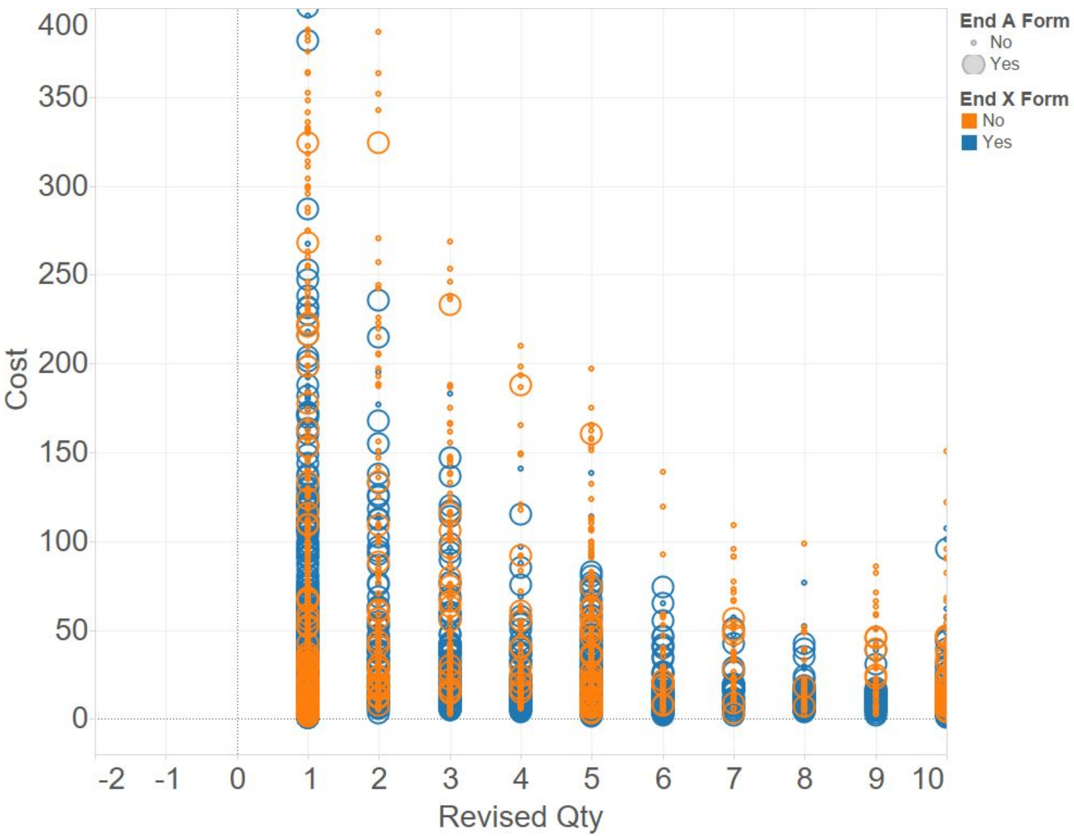
TA Bill of Materials
TA ID
Component 1-8 ID
Qty each component

Component specs
Component ID
Spec ID

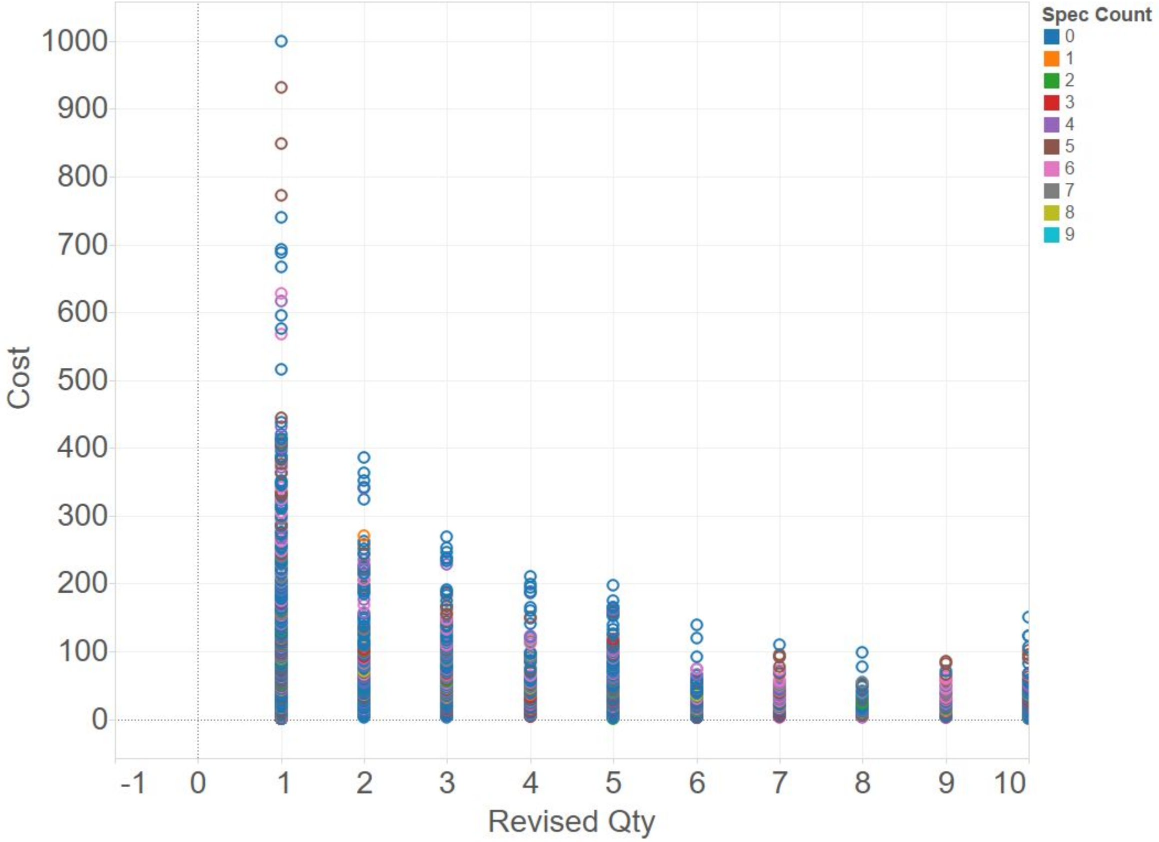
End fitting
End form ID
End forming (Y/N)

Component type
Component ID
Component descr.
Component Type ID

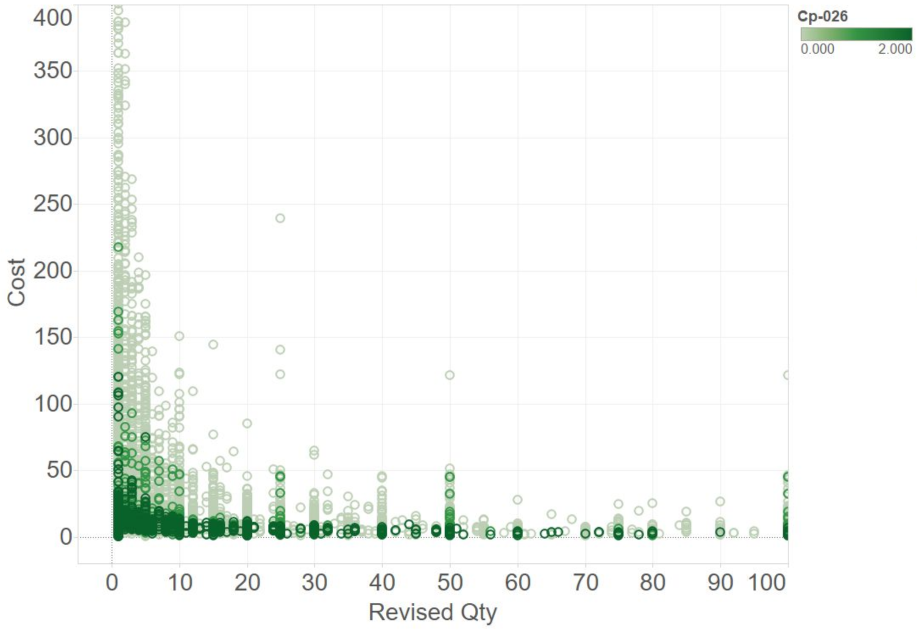
End forming doesn't seem to affect price



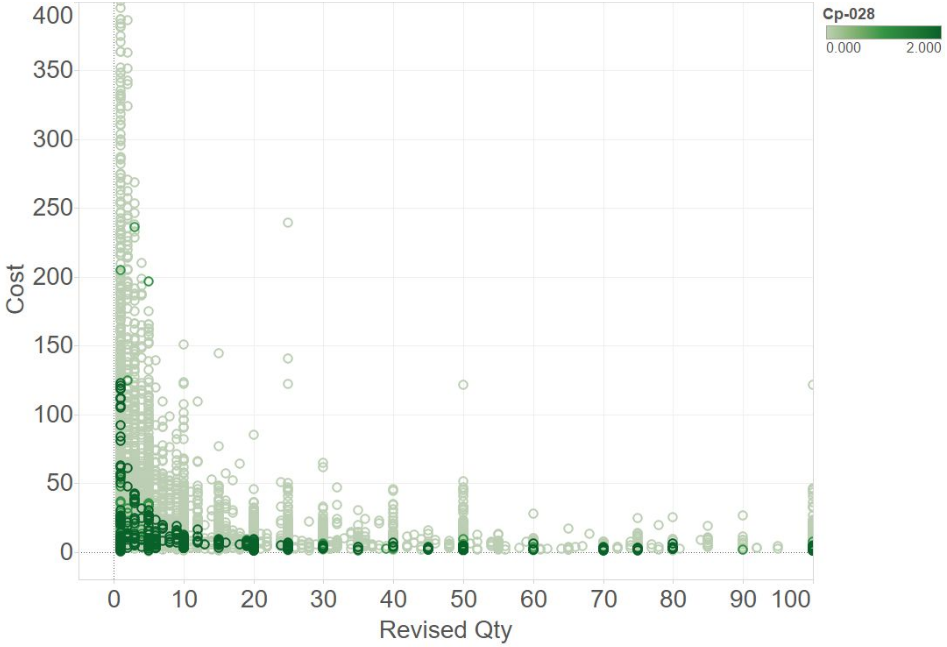
Number of specs doesn't seem to affect price



The number of components of certain types seems to be strongly correlated with price

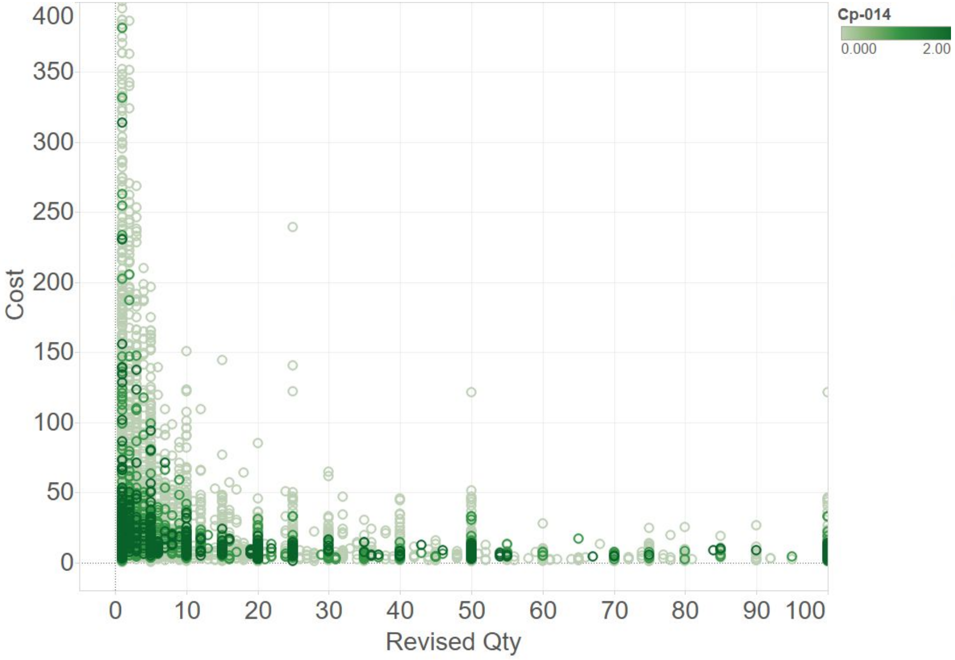


CP-026: JIC 37-45 NUT

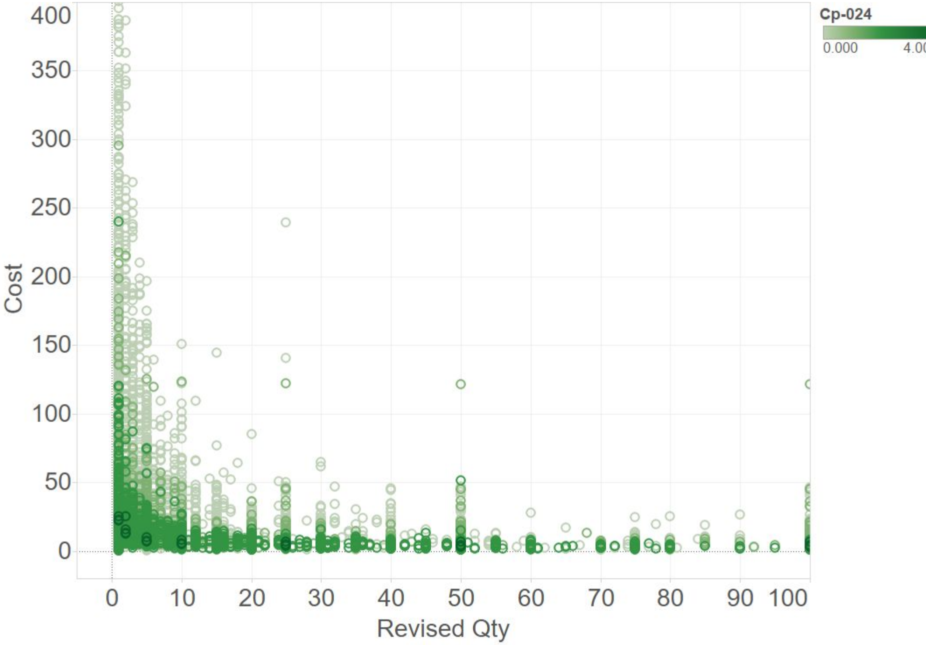


CP-028: Straight Adapter

The number of components of certain types seems to be strongly correlated with price



CP-014: Threaded straight



CP-024: Sleeve

Other data cleaning, #2

Data prep for modeling

Feature/Issue	Action	Data reduction
End A and X forming (Y/N)	Change to dummy variables	--
Component type for each component	Change to dummy variables	--

Modeling, #2

Models

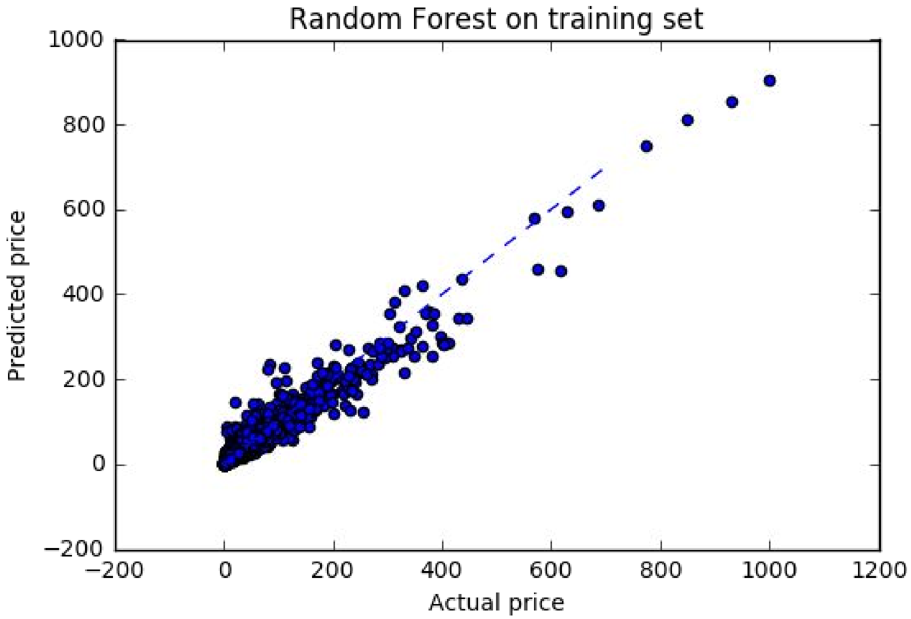
- Only Random Forest and Gradient Boosting were explored in this round

- No feature selection/dimensionality reduction, model tuning or log transformations were done. These, along with feature importances, will be explored during model optimization

- All features from round 1, and the additional features discussed earlier were included

(Base) Random Forest

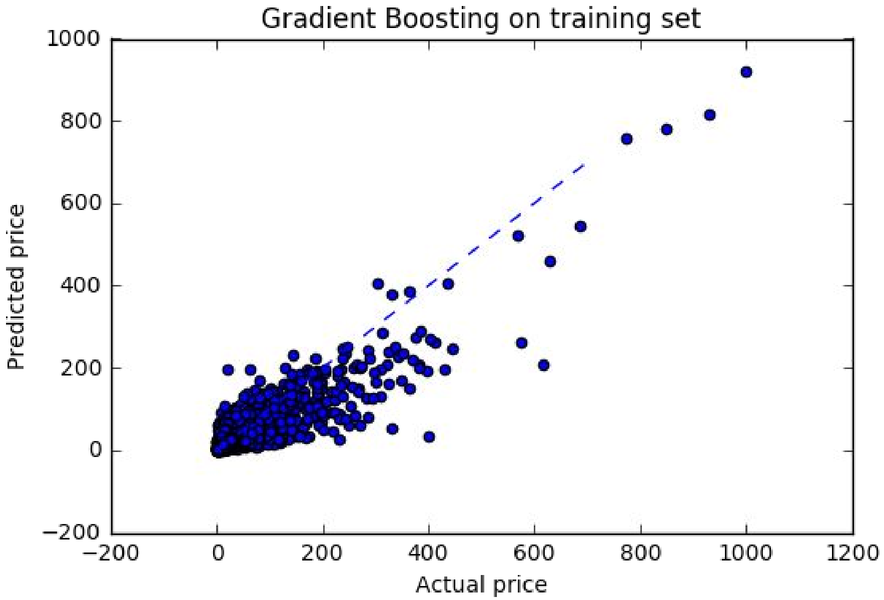
	R2
Training set (round 2, with add'l features)	0.951
Training set (round 1)	0.941
CV 5-fold (round 2 with add'l features)	0.707 +/- 0.029
CV 5-fold (round 1)	0.631 +/- 0.057



The additional features improved the CV score, but the base Random Forest model still suffers from a variance problem

(Base) Gradient Boosting

	R2
Training set (round 2, with add'l features)	0.779
Training set (round 1)	0.699
CV 5-fold (round 2 with add'l features)	0.649 +/- 0.041
CV 5-fold (round 1)	0.577 +/- 0.078



While performance has improved, the base Gradient Boosting model still lags the base Random Forest model. It continues to suffer from both bias and variance

Summary, #2

Key take-aways

- ❏ Random Forest model continues to perform better than Gradient Boosting
- ❏ The additional features improved the CV score for Random Forest … the additional features were important!
- ❏ Optimization of the Random Forest model (with the additional features) may help improve performance (on unseen data) even further

Model optimization

Options for tuning (variance reduction)

1. PCA for dimensionality reduction (prior to model fitting)

2. Hyperparameter search using GridSearchCV

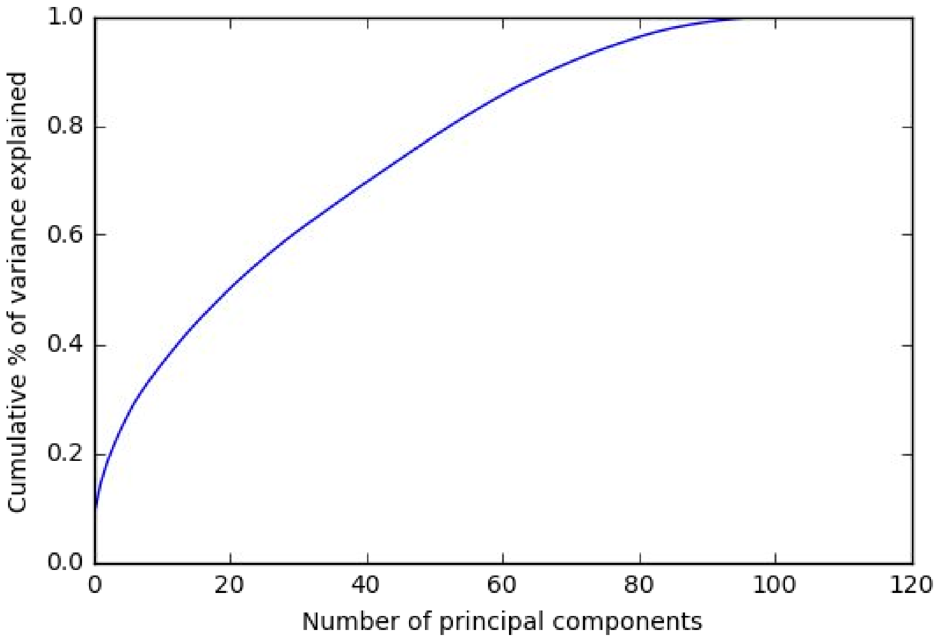
3. *Look for additional features that matter!*

Models and features

- The following models were explored
  - Random Forest
  - Extreme Gradient Boosting (aka regularized gradient boosting)

- The target was log transformed (new_target = log(cost+1)), in light of the relationship seen in EDA #1. The revised_qty was log transformed as well

1. PCA



- There is no single (linear combination of) feature that explains a large chunk of the variance
- **End forming** and **type of end fitting** (on ends “A” and “X”) contribute the most to the principal components that explain the most variance (i.e. have the highest eigenvalues)

90 principal components (out of 119 total) explain ~99% of the variance in the features

1. (Base) Random Forest, with and without PCA

	PCA (n_components)	R2
Training set	--	0.971
Training set	90	0.957
Training set	75	0.956
CV 5-fold	--	0.833 +/- 0.007**
CV 5-fold	90	0.749 +/- 0.005
CV 5-fold	75	0.748 +/- 0.007

There is more variance in the models with PCA … linear combinations of features may lose information in an inherently non-linear model

***** Increase from 0.707 in previous round attributed to use of log transformation of cost as target***

2. Random Forest GridSearch CV (on EC2)

	Base Model	GridSearchCV
R2 training	0.971	--
R2 CV	0.833 +/- 0.007	0.838 (Best score)
RMSLE test	--	0.364

- ExtraTrees Regressor model performed similar to Random Forest
- Bagging Regressor (which enables the use of subsamples to reduce variance) also gave R2 CV of ~0.84

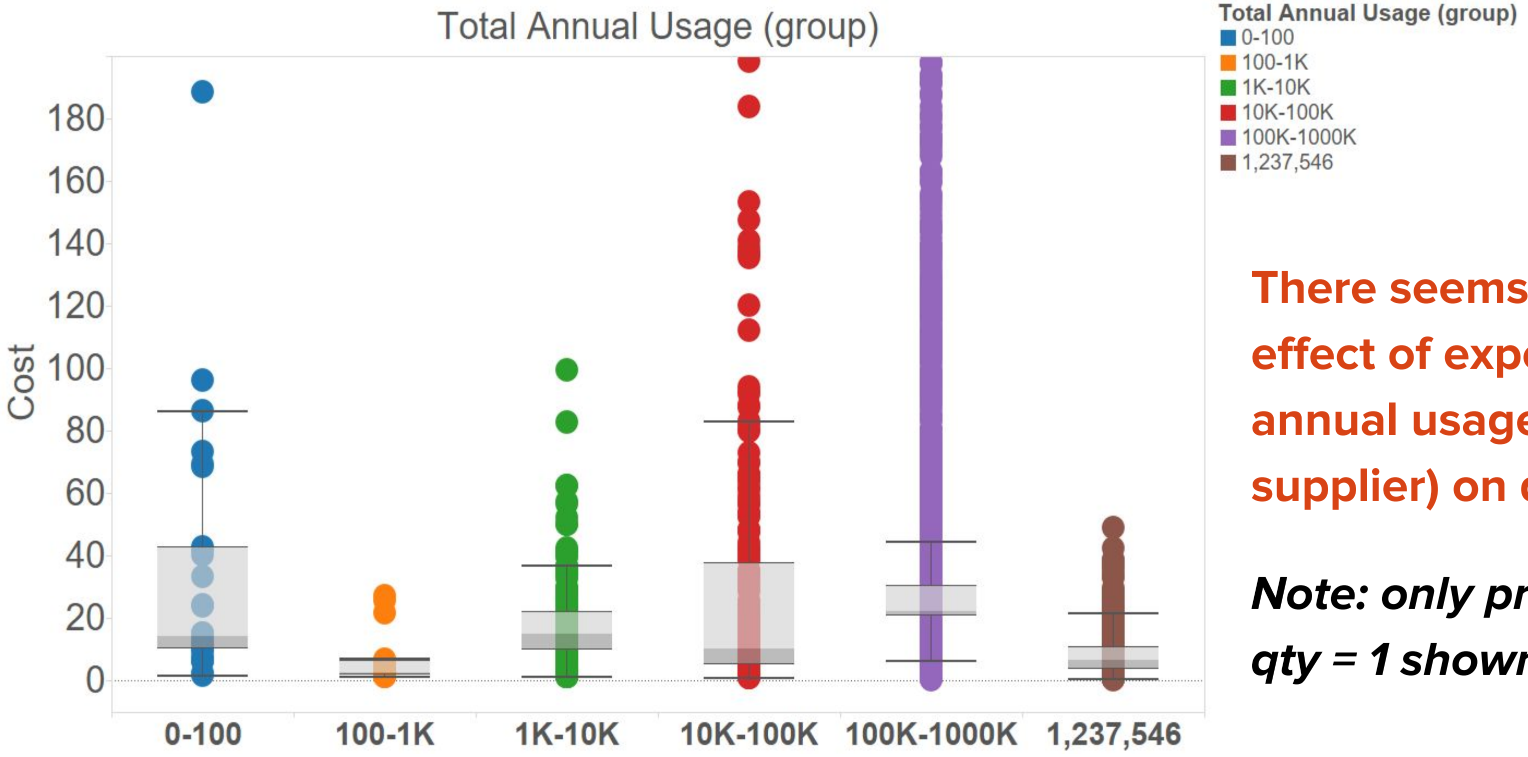
Random Forest still suffers from a variance problem

2. Extreme Gradient Boosting (xgboost)

- ❏ R2 CV (best score): 0.86
- ❏ RMSLE test: 0.348

While xgboost is slightly better than Random Forest, it is still overfitting. We need to search for additional features (from our original hypotheses) that matter!

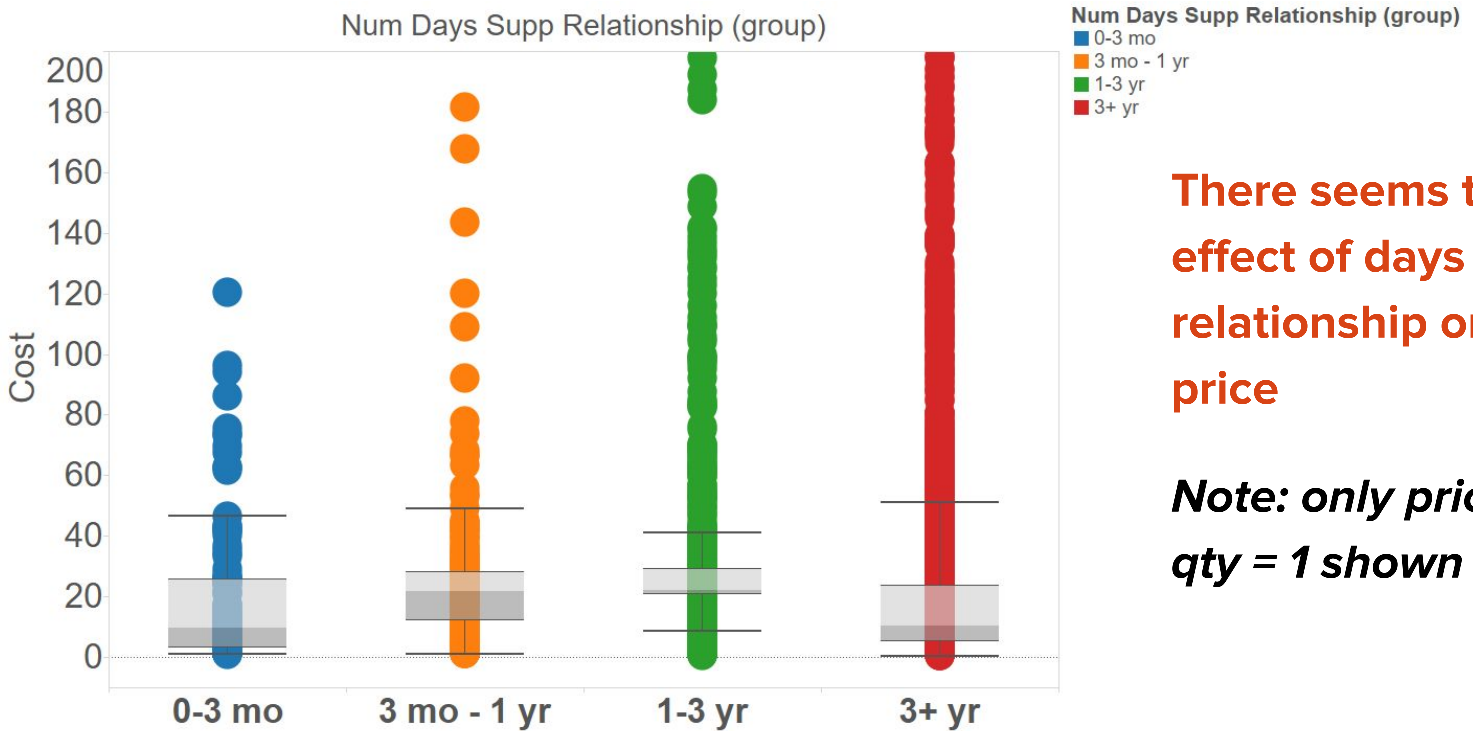
3. Additional features - total annual usage by supplier



There seems to be some effect of expected total annual usage (for each supplier) on quoted price

Note: only prices for TA qty = 1 shown for clarity

3. Additional features - days of relationship of supplier with Caterpillar at time of quote



There seems to be some effect of days of relationship on quoted price

Note: only prices for TA qty = 1 shown for clarity

Model performance - with new features (unscaled)

Note: Unscaled features were used to ensure that non-linearities in the model don't get "smoothed" out by scaling

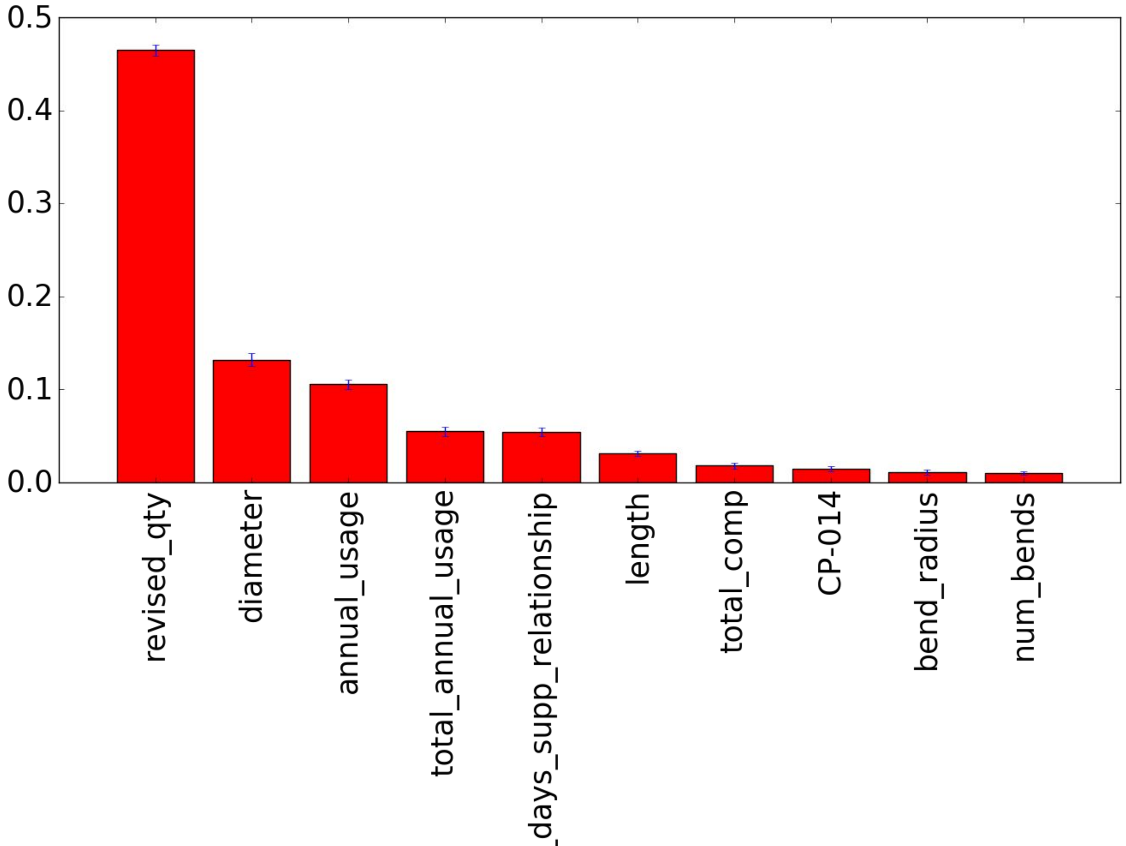
	Random Forest	xgboost
R2 CV(Best score)	0.886	0.914
R2 test	0.901	0.934
RMSLE test	**0.26**	**0.212**

Much closer to the upper bound of performance than lower bound … the model performs fairly well!

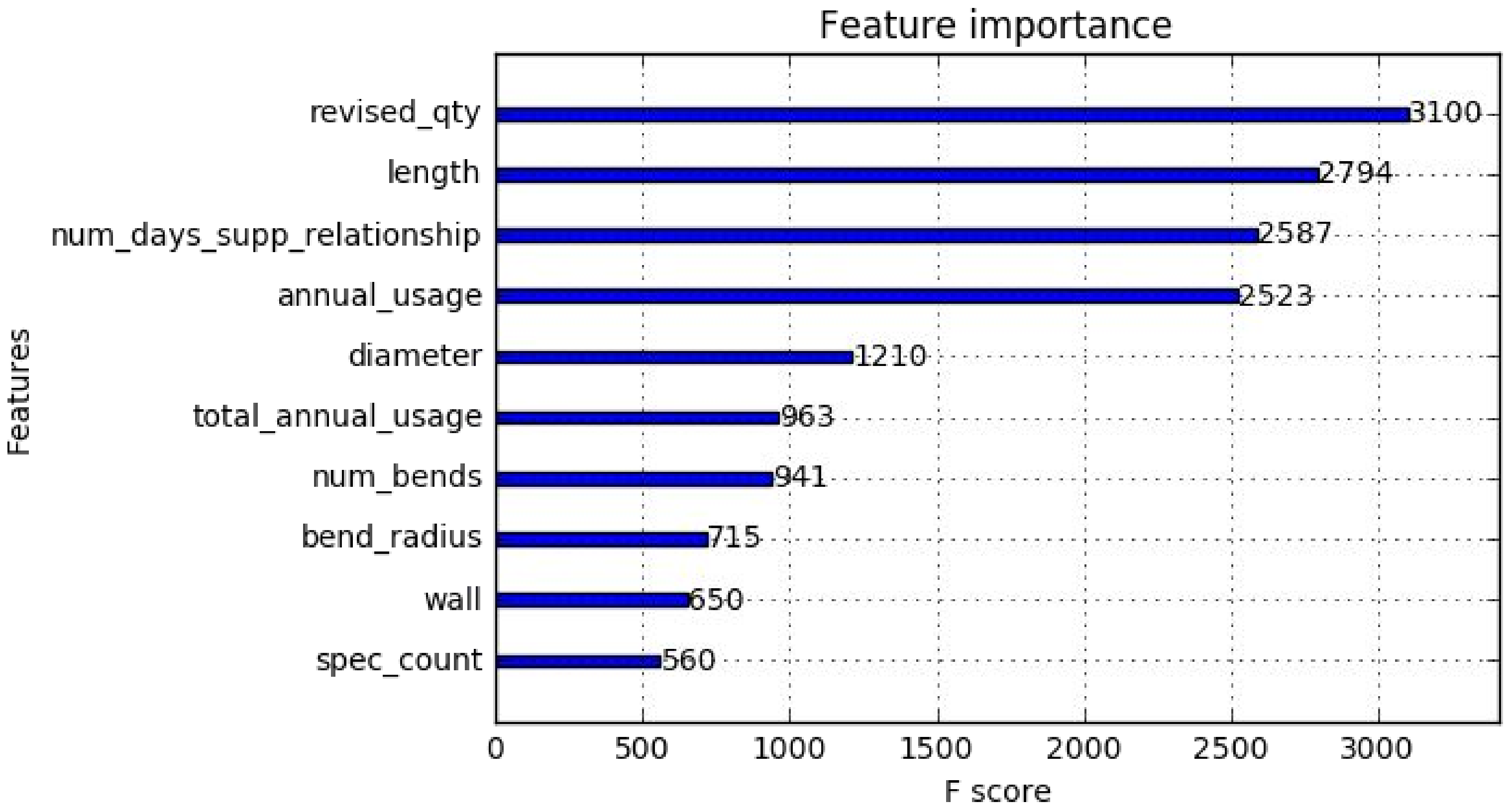
Winning Kaggle score (**<u>on Kaggle's test set</u>**): 0.197. **Potential top 3% finish**

Scores are much improved with the new features, and using unscaled data

Feature importances (Random Forest)



Feature importances (xgboost)



Summary

1. Quoted price is a non-linear function of features
2. As expected, order quantity is the most significant factor affecting price
3. TA diameter and length are important predictors of price, likely as a proxy for amount of material used. Amount material α dia^2 x length
4. Annual usage (per TA), total annual usage (per supplier) and days of supplier relationship strongly affect price

Recommendations

1. Most TAs seem to be sole-sourced. An obvious area of focus may be to qualify/help bring on board additional suppliers to increase competition, esp. for higher price TAs
2. An important area of focus for TA portfolio price reductions should be larger diameter TAs (esp. thick wall). Increasing annual procurement may help reduce prices
3. Increase use of CP-014 (threaded straight component) where possible to reduce price

Next steps

1. Given that TA diameter, TA length and total number of components per TA are important factors for price, it may be interesting to see if factors such as TA volume and total weight (which are more directly related to amount of material) improve model performance
2. Given that materials of construction or number of specifications per TA don’t play a big role in determining price, this may present an opportunity to change these if required by the technical team