Techlent Regression Challenge

Hong Tang

Summary and Recommendation

- **Why:** Predict the price make-up product and provide insights explain recommended price to customers
- **Method: End to End regression exercises**
- **Learning and Challenges**
  - Realistic workflow for data cleaning, EDA, Feature engineering
  - Price and cost are two leading indicators for price prediction
  - Strong collinearity among cost and weight, depth and height (Potential overfitting)
  - Significant improvement of model performance from Random Forest Model to XGBOOST model
  - Many interesting patterns in time series data such as most transaction happens on Friday; there are two high price seasons etc. We could use these trends to improve model prediction later
- **Lessons Learned and Best Practices**
  - Focus on feature engineering, and data QC
  - ML Pipeline avoid data leakage, makes modeling easy to read and maintain
  - XGBOOST outperforms Random Forest in the preliminary test
- **Plan Forward:**
- Improve feature engineering;

ML Process

Cycle 1

Feature EDA

Existing numerical features

Model Selection

Simple Linear regression

Cycle 2

Feature EDA

Time features

Numerical features
Categorical features

More data cleaning

Joint resample Data decisions

Modeling H.P. tunning

RandomForest Model

XGboost

Pipeline

Cycle 3

Future feature engineering

Time Series Analysis

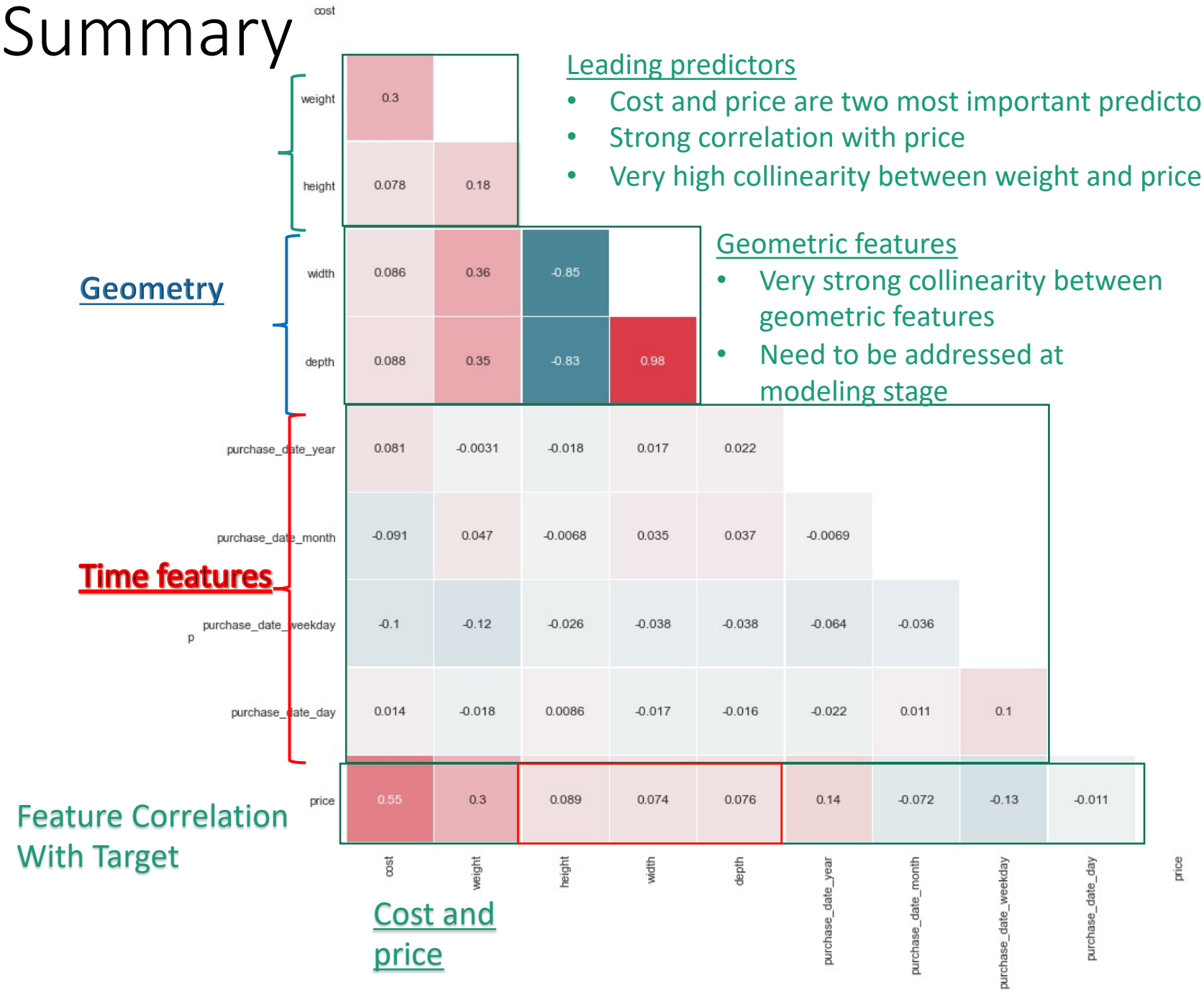
Different Cost function focus on residual trends

Model Deployment

Summary

recommendation

Summary



Geometry

Leading predictors

- Cost and price are two most important predictor
- Strong correlation with price
- Very high collinearity between weight and price

Geometric features

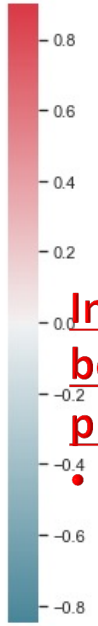
- Very strong collinearity between geometric features
- Need to be addressed at modeling stage

Time features

Feature Correlation With Target

Cost and price

Note: product type and ingredient and product levels are not included in this correlation matrix

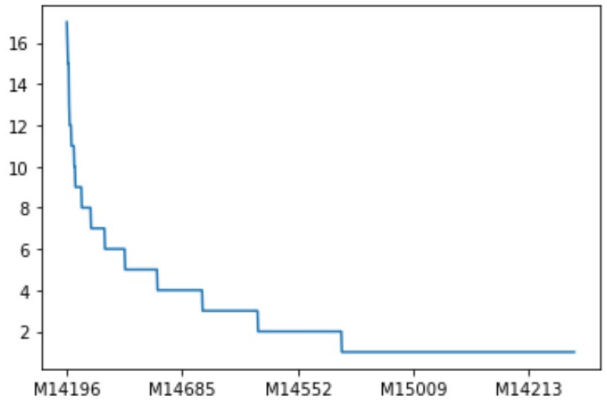


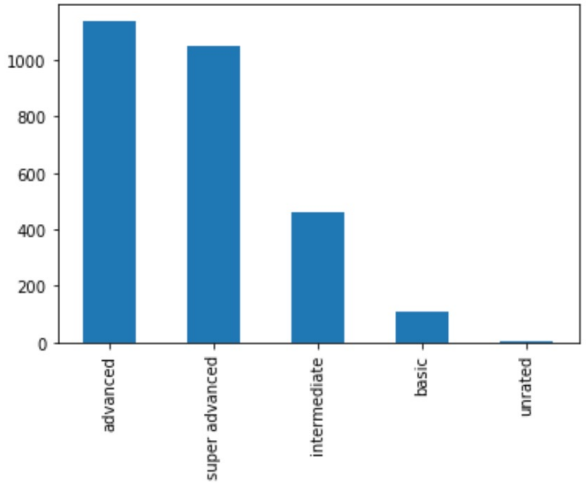
Interesting correlation between time features and price

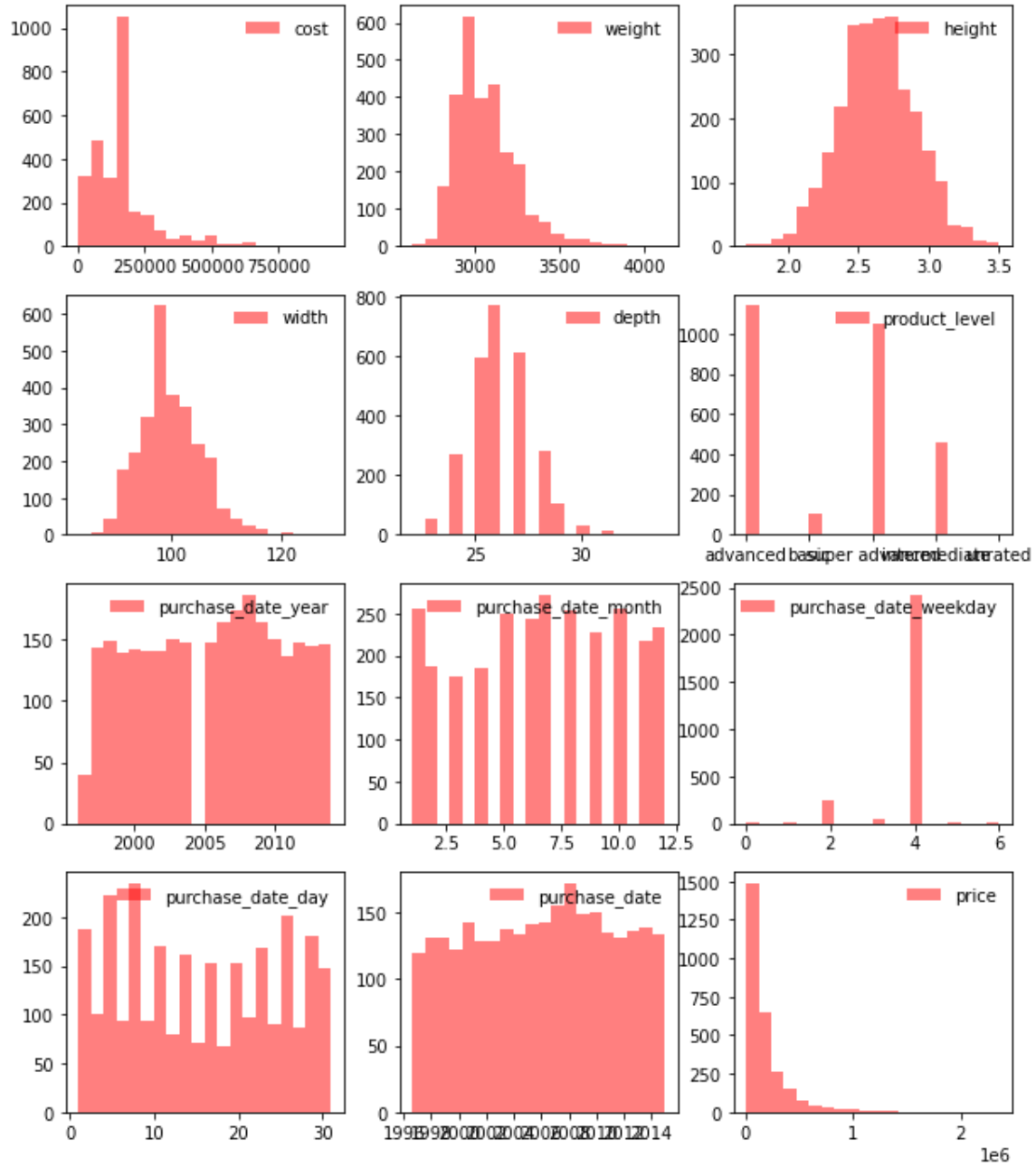
- **Later years the produt price becomes more expensive due to increase cost**
- **Friday has most transactions**
- **Basic product price is differe**
- **More feature engineering could be done in time features(seasonality etc.)**

Feature EDA

- Cost and weight has similar distribution as price
- Most transactions happens on Friday
- Few main makers, frequency transformer is used
- A gap around 2004- 2005(data quality issue?)
- Basic product quantity is much less than other product level
- Ingridient has skewed distribution

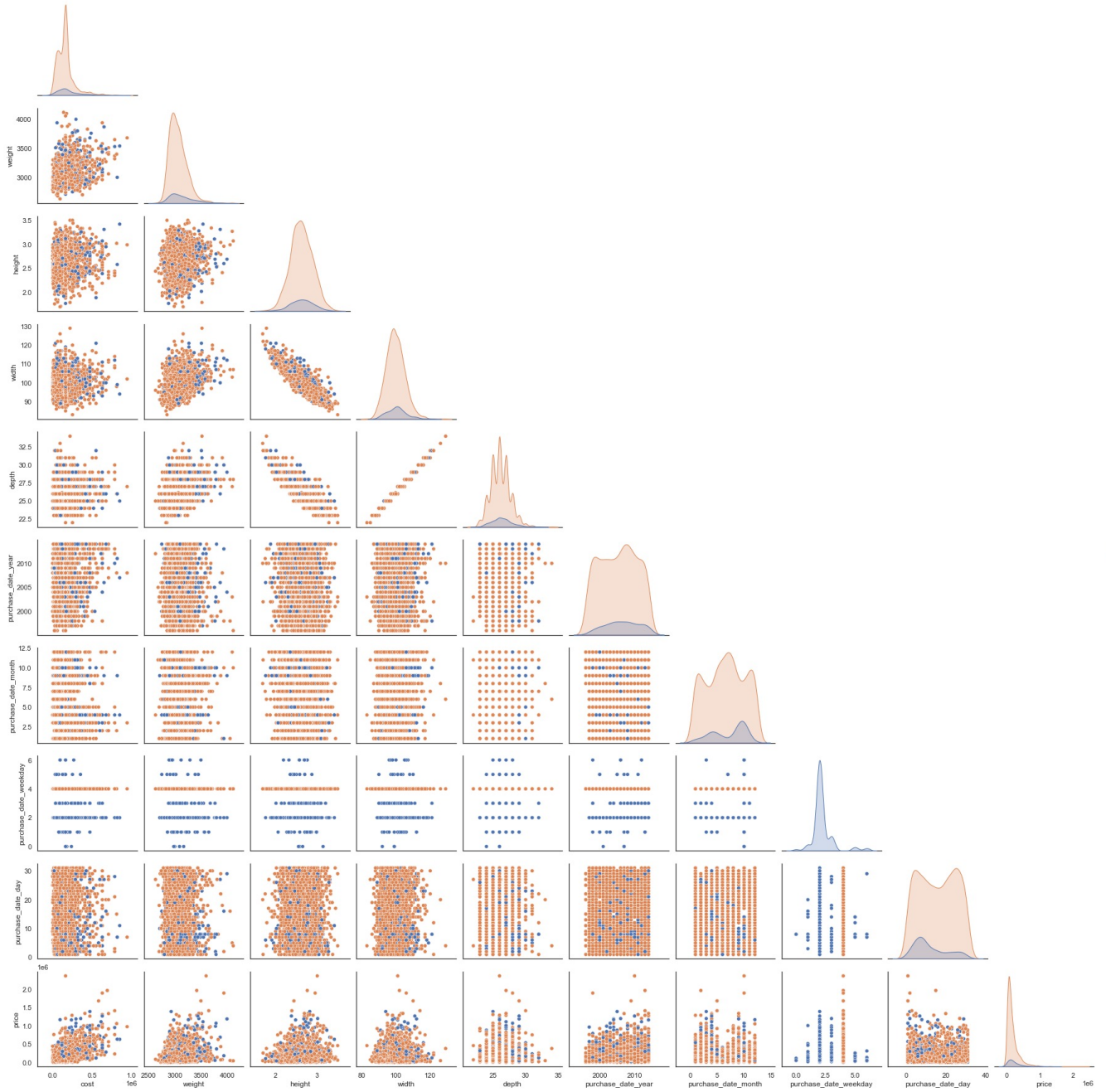




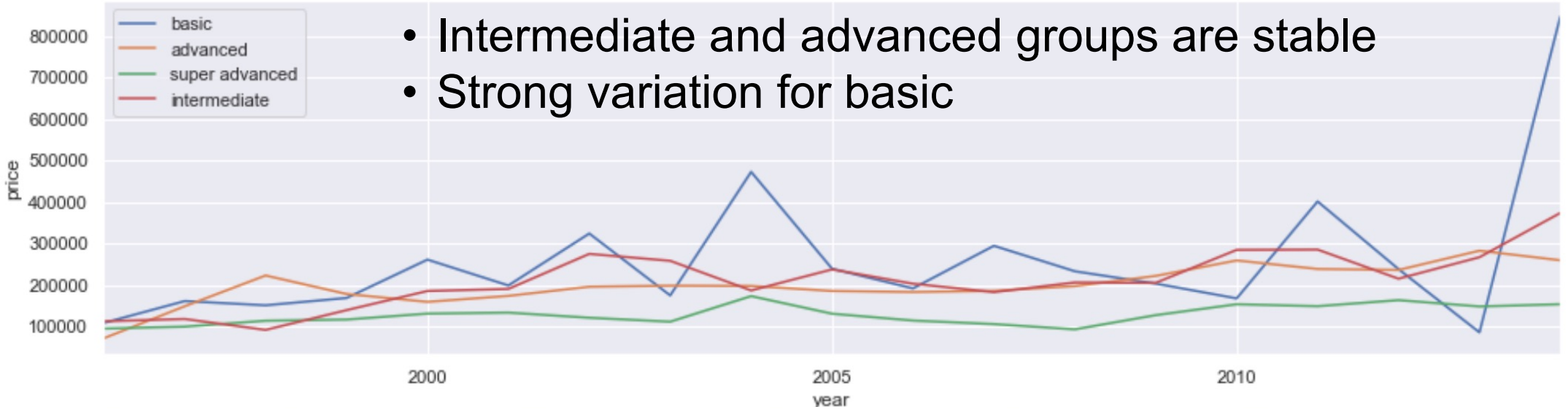


Feature EDA

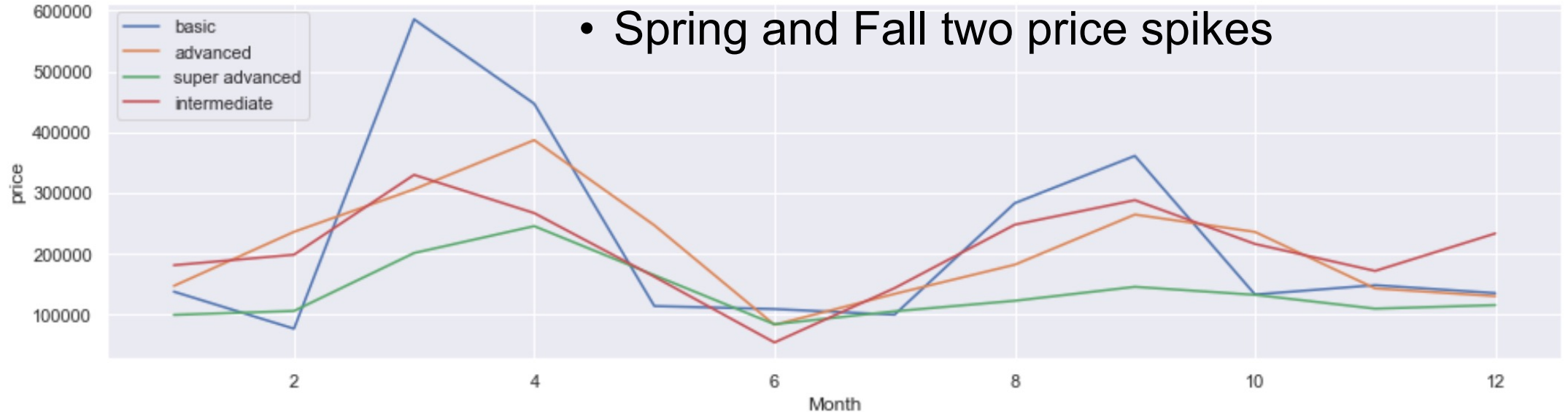
- Created Flag of Friday-no-Friday
- There is no clear pattern Friday is a good predictor of price
- Verifies the colinearlity among features



Time features



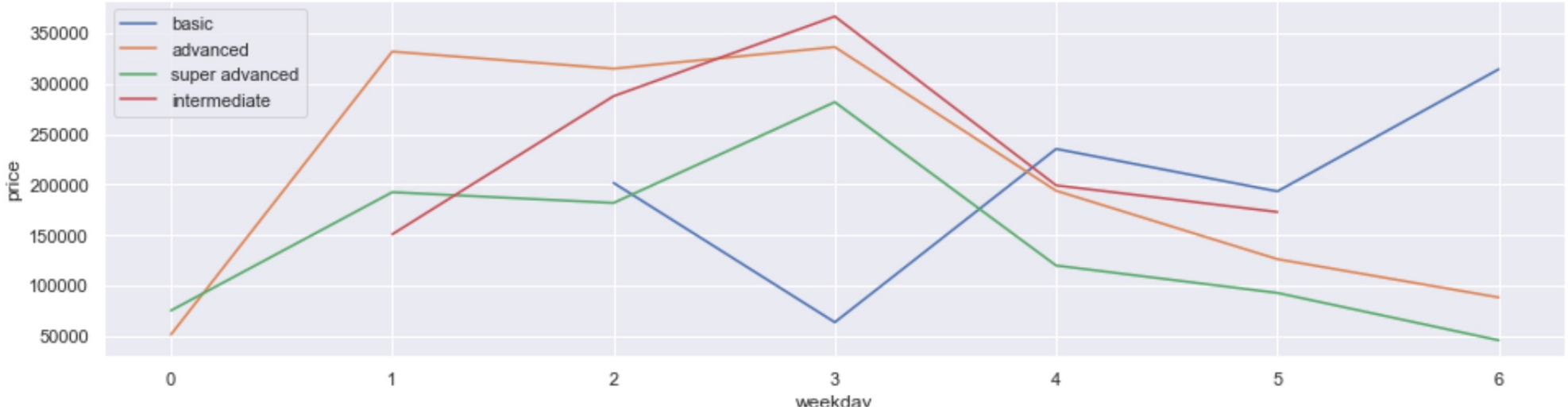
- Intermediate and advanced groups are stable
- Strong variation for basic



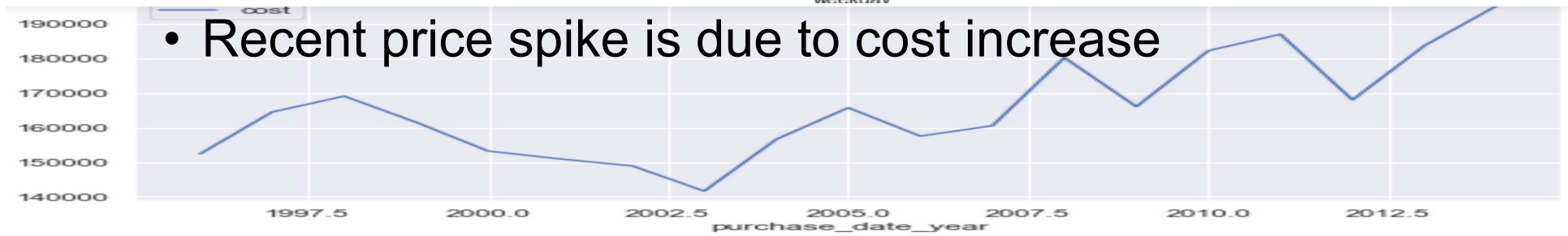
- Spring and Fall two price spikes

Time features

- Basic and other brands has different trend
- Might relate to different customer groups: working mom vs house wives (no gender bias)?



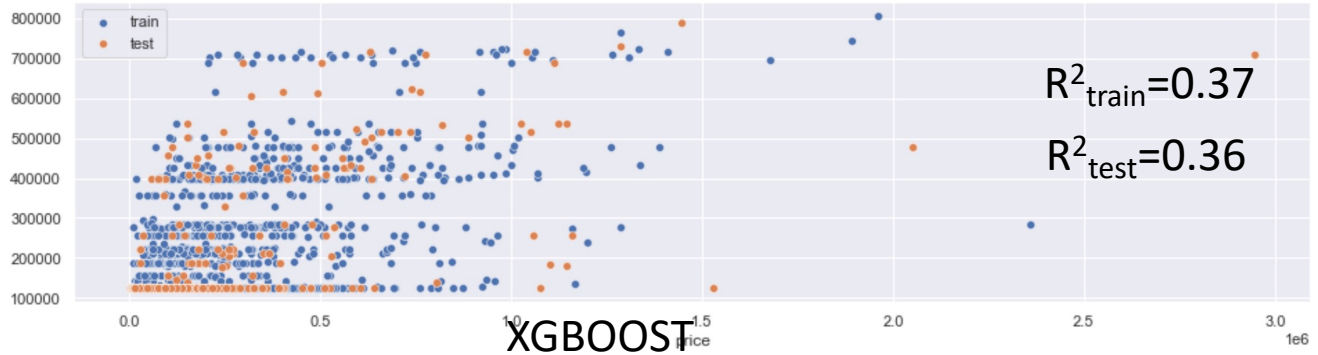
- Recent price spike is due to cost increase



Model improvement

- R^2 improved from 0.36 to 0.43 train and test
- Future work
  - Extracting and include more time features into modeling
  - Overfitting in high price range; try to improve by further segmentation

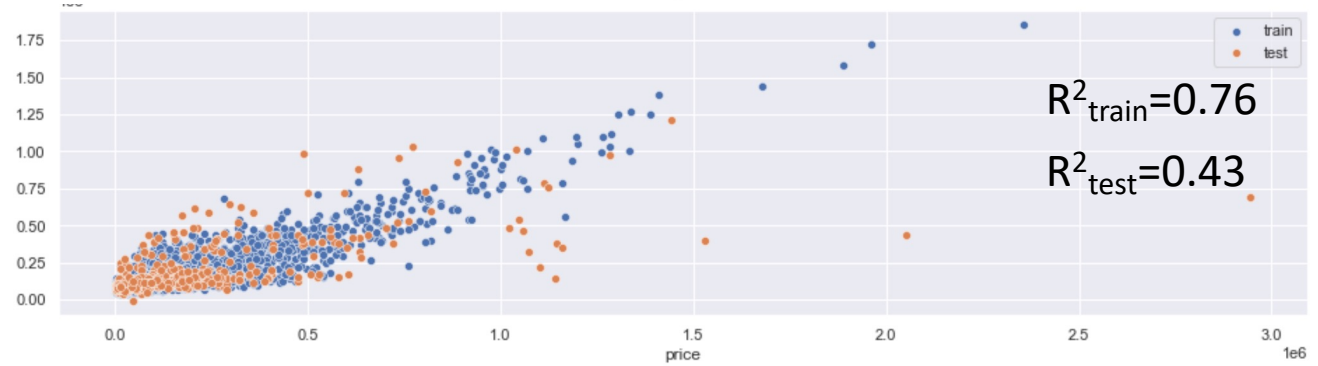
Random Forest



R^2_{train}=0.37

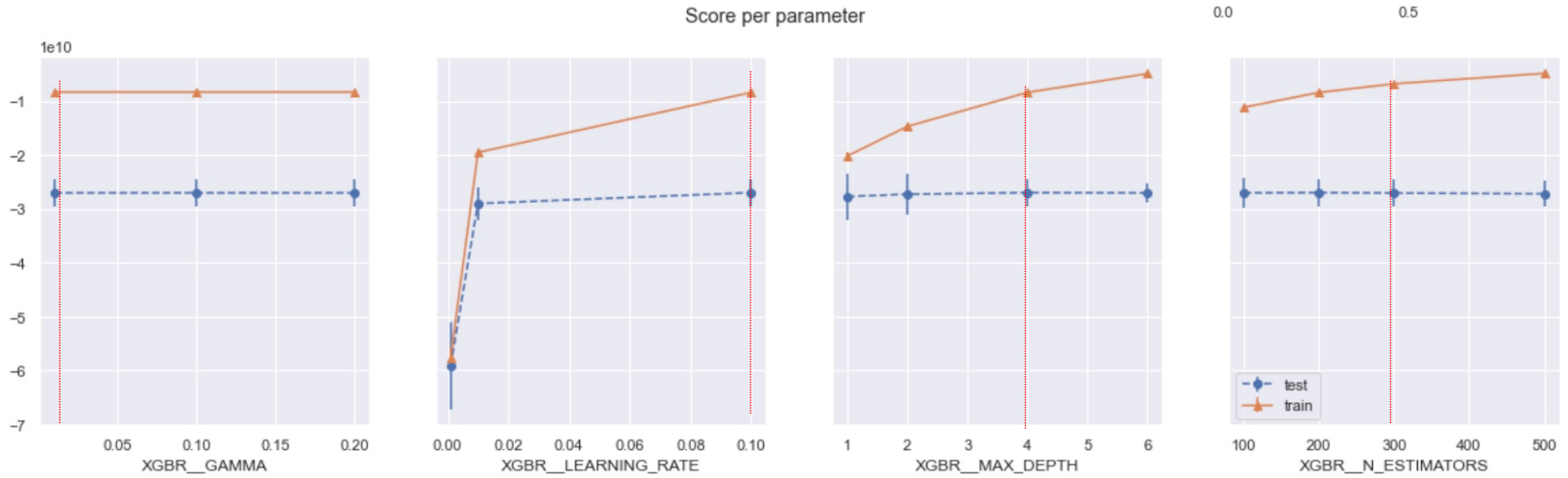
R^2_{test}=0.36

XGBOOST



R^2_{train}=0.76

R^2_{test}=0.43



XGBOOST Grid Search CV
Derived Best Estimators