Big Data is **also** Big Compute
Better Price Forecasting using Machine Learning

Avishkar Misra PhD – avishkar.misra@oracle.com
Big Data & Analytics Platform Team

ORACLE

Abstract

Oracle Big Data Appliance and Oracle Big Data Cloud Service offer a fantastic computation platform to scale up experimentation and speed up time to insight. The Oracle Big Data & Analytics Platform team, worked with a leading North American commodity producer in 2016, to help improve the granularity and accuracy of their prediction for the price of the commodity. An accurate price forecast model has a potential benefit of $25.6M over a 3 year period for the commodity producer. We used Python and PySpark running on Oracle Big Data Appliance to try out 39,936 modeling combinations in a matter of minutes to find a better algorithm. The more accurate algorithm was able to predict the price within +/-5% of the actual price 73% of the time compared to 40% for the algorithm the customer had developed and fine-tuned over many years. Oracle's Big Data platforms offer a way to dramatically scale the number of experiments, helping the scientists and analysts quickly narrow their focus on the most relevant and effective techniques to impact their business's bottom line.

Blog Post: https://a-misra.com/2016/08/01/big-data-is-also-big-compute/

ORACLE®

Big Data & Data Science Advisory Services

Customer Engagements that bring industry experience and best practices to demonstrate and prototype solutions for customer in line with their strategic business goals.

1 Business Value Definition

2 Solution Architecture

3 Data Science & Machine Learning
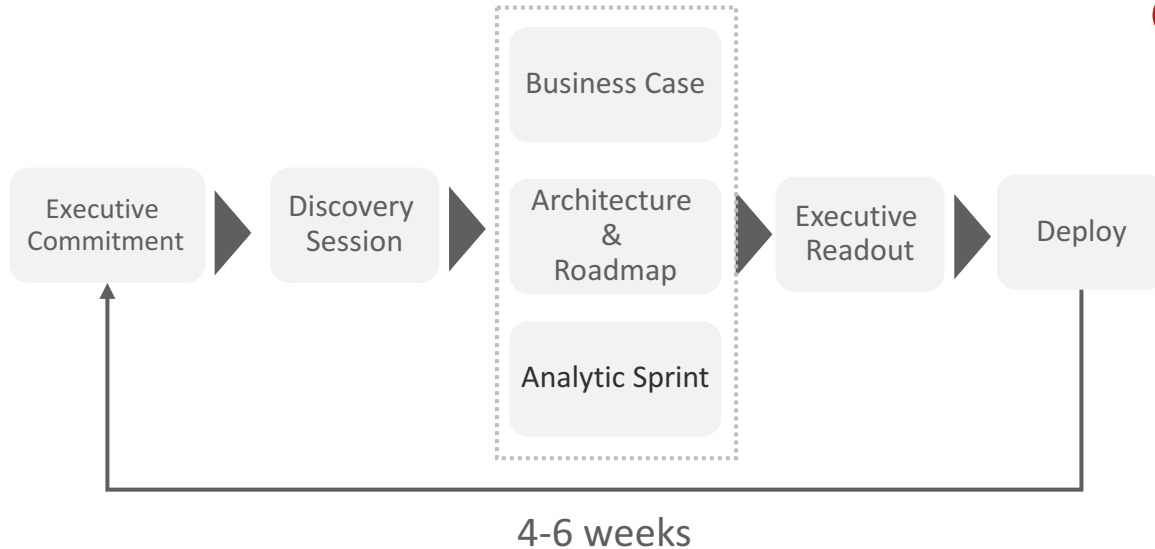
4 Analytics Design Hub

Team skill-sets:
- Data Scientists
- Data Wranglers
- Architects
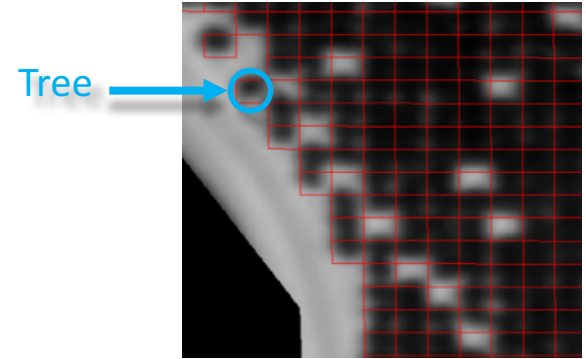- Business Analysts

Team backgrounds from:
- Amazon
- MSFT
- Deloitte
- IBM
- Teradata
- etc.

ORACLE

Big Data Analytics Sprint

Define, design, develop and demonstrate the value of a solution quickly.

```
Executive      →   Discovery   →   ┌ Business Case ┐   →   Executive   →   Deploy
Commitment         Session         │               │       Readout
                                   │ Architecture  │
                                   │ & Roadmap     │
                                   │               │
                                   └ Analytic Sprint┘
```

4-6 weeks

1 **Agriculture**: aerial image analysis

Tree →

2 **Utilities**: fraud detection

ORACLE®

Today: Commodity Price Forecasting

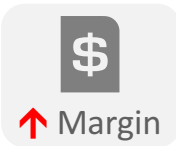A large North American Lumber producer looking to improve weekly price forecasts

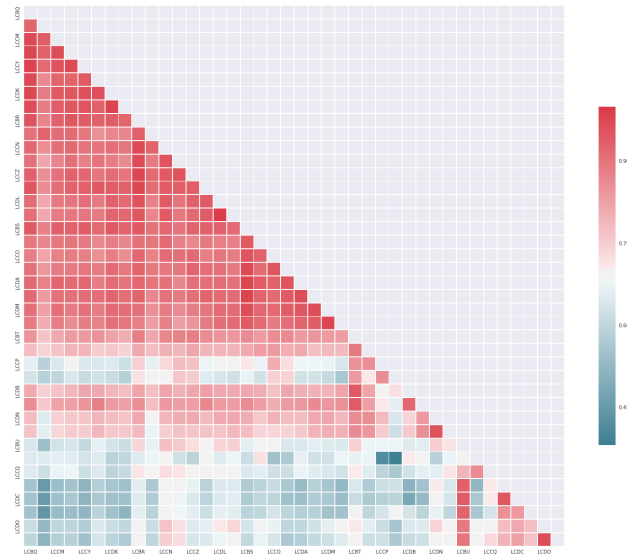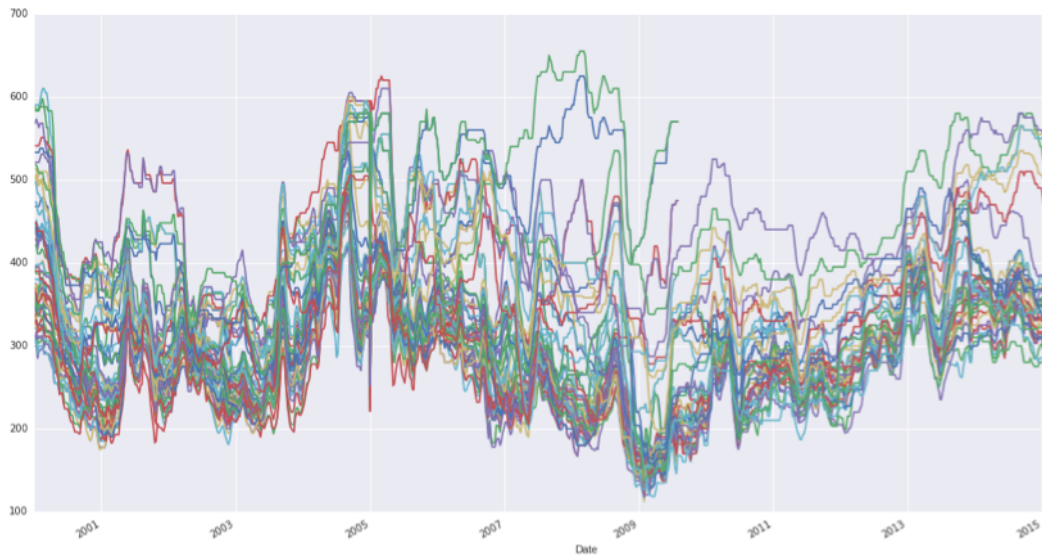Spot market traders speculated on weekly commodity price using:

- prior week's market price
- 90-day forecast – Tuned over 10 years
- gut instinct

Decide: **sell** vs **hold** inventory

More accurate price prediction can

$ ↑ Margin

Lumber lumber everywhere

LCBQ - KD Western Spruce-Pine-Fir #2&Btr 2x4 random.

Common benchmark, since it has high correlation with other types of lumber.

What factors influence the price?

Current Price	Weekly (every Friday)
Economic Activity in Construction & Building	Monthly (3-month lag)
Currency Exchange Rates	Daily
Weather *new	Hourly

Normalize the data for weekly predictions by

- over-sampling

- aggregation

- sub-sampling

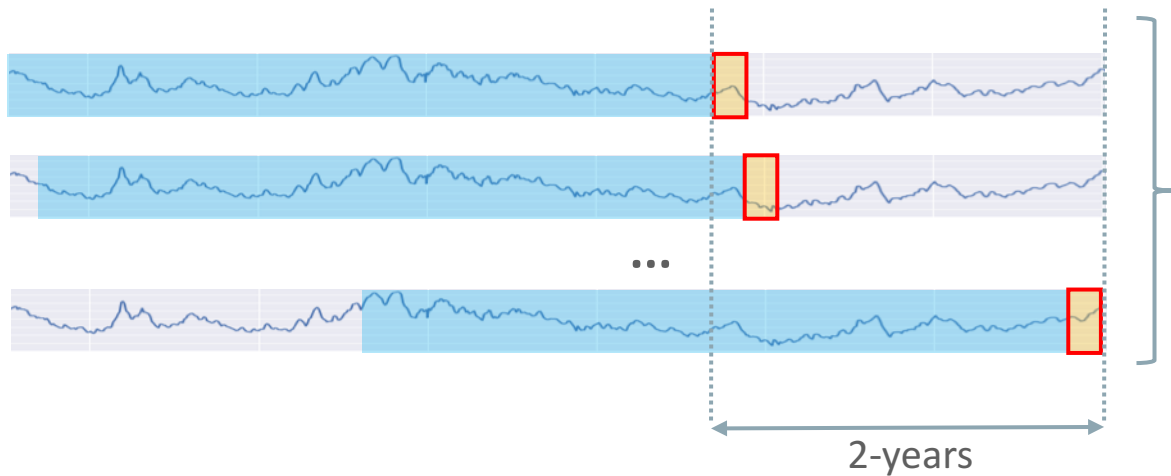ORACLE®

Weather, but for where?
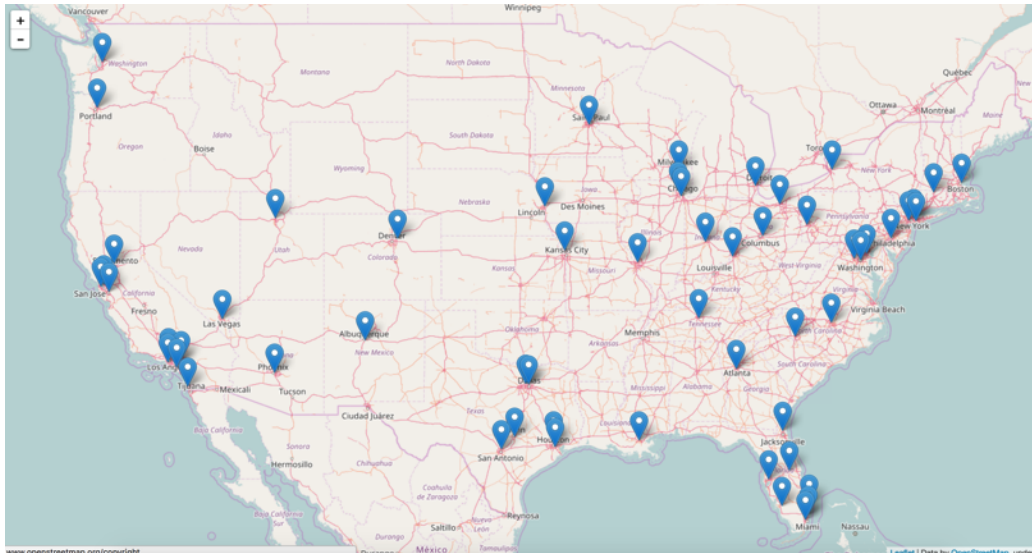
Focus: 59 major population centers – where people live and likely to renovate + build.

Weather Data:

- Weekly Total Precipitation

- Weekly Average Temperature

Evaluation Methodology

Time-series prediction, so cannot use traditional k-fold validation.

For 104 weeks:

- Train models for each week.

- Predict for next week.

- Evaluate prediction accuracy

2-years

ORACLE

Q: Machine Learning pipeline options – Which do we pick?

|		
Feature Generation	None, Rolling Window {7, 13, 26, 52}, Polynomial Interactions, All	6
Feature Normalizer	None, Min-Max	2
Feature Selector	None, PCA	2
Modeling Objective	Classifier, Regression	2

Modeling Algorithm	Logistic Regression Linear Regression
	Decision Trees Ridge Regression
	Extra Trees Lasso Regression
	Ada Boost Decision Tree Regressor
	Gradient Boosting Ada Boost Tree Regressor
	Random Forest Extra Tree Regressor
	     Gradient Boosting Regressor
	     Random Forest Regressor
		6 | 8

ORACLE

Answer: Experiment!

If you double the number of experiments you do per year you're going to double your inventiveness.

– Jeff Bezos

Data Scientists

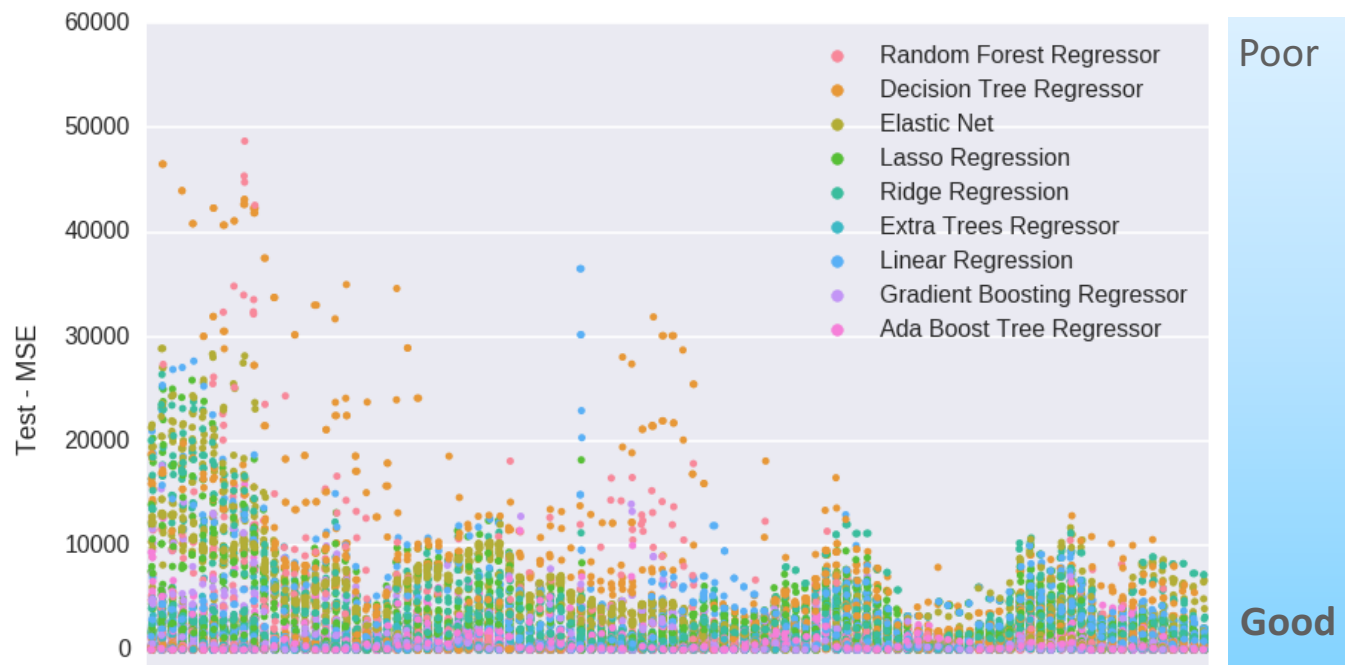Oracle **Big Data** Cloud Service

ORACLE

Classification – Can we predict if the price will go up?

Trained and evaluated - 14,976 Classifiers
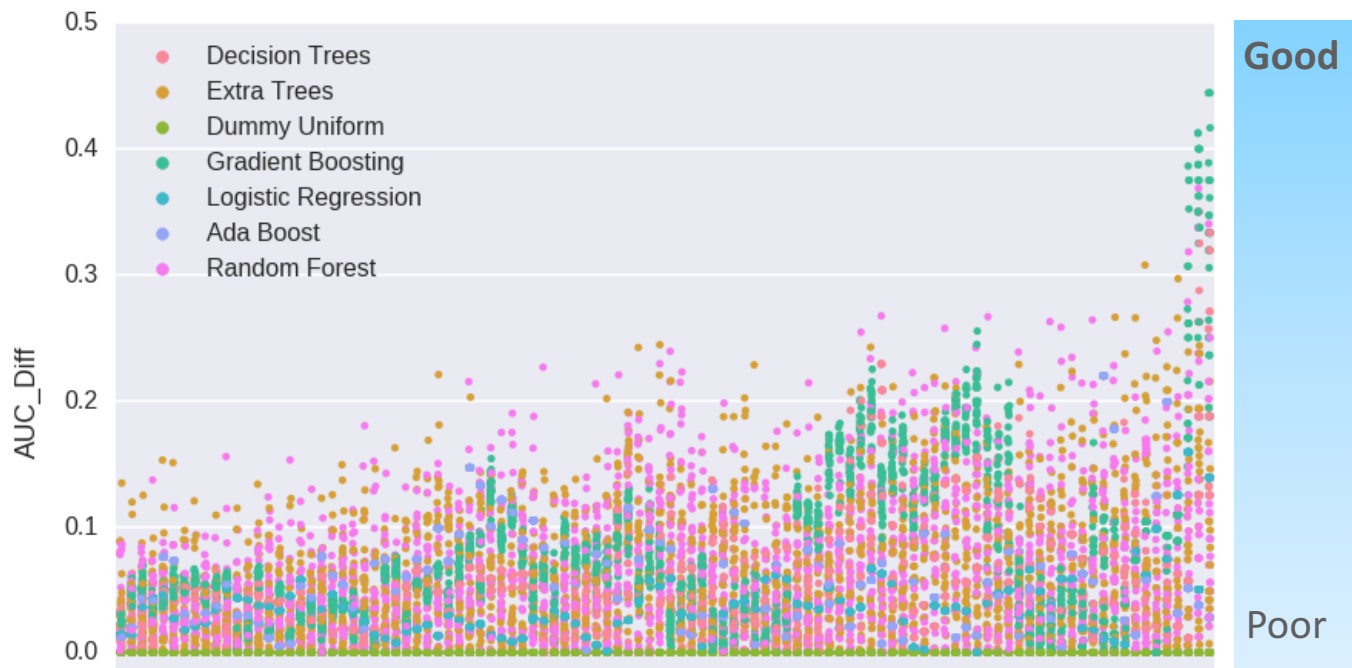
Regressors – Can we predict the price?

Trained and evaluated - 22,464 Regressors

Takeaways

Evaluated **37,440** models in less than 15mins!!

Classifiers:
- Ensemble are better - Extra Trees, Gradient Boosting, Random Forest, Ada Boost.
- Min-Max normalization of feature values not useful
- Principal Component Analysis very helpful

Regressors:
- Lot of model over-fitting!
- Extra Tree Regressors work very well
- Principal Component Analysis not helpful

Insights → A better price prediction algorithm
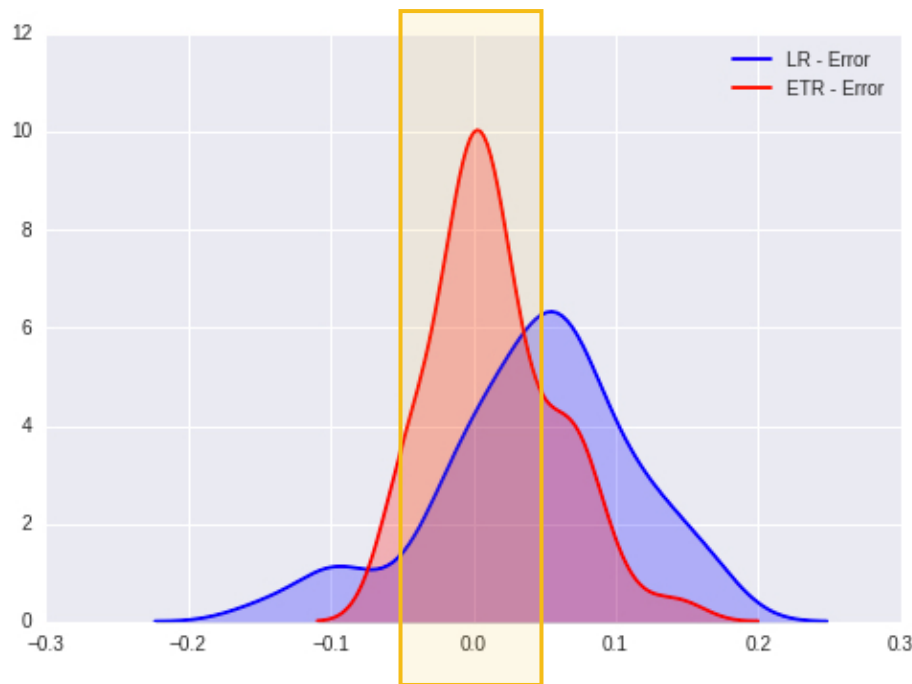
Target is Actual Price +/- 5%

Linear Regression: **40%** of time in target.

Extra-Tree Regression: **73%** of time in target.

$ ↑ **$19.4m** to **$25.6m** margin over 3 years

ORACLE

Price Forecasting for Commodity Goods

Objective: Improve price forecasting on finished commodity goods to optimize margin dollars across all channels and enhance customer upselling capabilities for traders

Business Challenge / Opportunity

- Sales Reps (spot market traders) relied on prior week market pricing and long-range (90) data forecasts to make a 'gut instinct' decision – weekly price forecasts did not exist

- Business Analysts manually pulled information from a variety of sources into an Excel workbook – often having to retype market trend information from PDFs

- While the current process was deemed helpful for long range strategic forecasting, traders needed short-term forecasts in order to better price goods and determine whether to sell or hold inventory on a weekly basis

Example

Leading North American Lumber Producer for Builders and Home Improvement Centers

Analytic Sprint Results

- Demonstrated the inclusion of external datasets such as **weather** into the mix of data to improve forecast accuracy

- Demonstrated new **machine learning** techniques and the ability to scale client experimentation with forecast modeling to **37k** analytic models in a matter of **minutes**

- Improved prediction **granularity** from monthly to weekly, and **accuracy** to within **+/- 5%** of the actual market price **73%** of the time, compared with **40%** for the algorithm used by the client

Proposed Solution

- Oracle Big Data Cloud Services, including: Cloudera Hadoop Dist., Spark, Jupyter, Connectors, Data Integrator, Oracle R Advanced Analytics for Hadoop, Spatial & Graph, Storage

Potential Benefits

- Over a 3 year horizon, potential incremental **gross margin** from improved pricing and reduced **operating costs** ranged from **$19.4m** to **$25.6m** in incremental benefits, and a **4,342%** ROI

Big Data & Data Science Advisory Services

Get in touch with us!

Email: avishkar.misra@oracle.com

Blog: http://a-misra.com

Oracle External: https://www.oracle.com/big-data

Oracle Internal: http://bdcoe.us.oracle.com

This work: https://**a-misra.com**/2016/08/01/**big-data-is-also-big-compute**/