Forecasting, Prediction Models, and Times Series Analysis with Oracle Business Intelligence and Analytics

Rittman Mead BI Forum 2013

Dan Vlamis and Tim Vlamis
Vlamis Software Solutions
816-781-2880
http://www.vlamis.com
Presentation Agenda

• Understanding classification and forecasting (predictions)
• Use of Geneva Forecasting engine in Oracle OLAP
  • Holt-Winters and time series
  • Parameter choices
• ARIMA forecasting algorithm in R
  • Use Oracle R Enterprise
• Use of time dimension and time series functions in OBI
Dan Vlamis and Vlamis Software Solutions

- Founded in 1992 by Dan Vlamis in Kansas City, MO
- Developed/implemented more than 200 Oracle BI systems
- Specializes in ORACLE-based:
  - Business Intelligence
  - Analytic Options to Oracle DB (OLAP, Data Mining, Spatial)
  - Data Warehousing
  - Training and mentoring
- Expert presenter at major Oracle conferences
- www.vlamis.com (blog, papers, newsletters, services)
- Co-authored book “Oracle Essbase & Oracle OLAP”
- Beta tester for OBIEE 11g
- Reseller for Simba and Nokia map data for OBIEE
- HOL Coordinator for BIWA Summit 2013
Tim Vlamis’ Bio

• 20+ years experience in business modeling and valuation, forecasting, and scenario analyses
• Trainer for Oracle University Two-Day Data Mining Course
• Professional Certified Marketer (PCM) from AMA
• Active Member of NICO (Northwestern Institute on Complex Systems)
• Adjunct Professor of Business Benedictine College
• MBA Kellogg School of Management (Northwestern)
• BA Economics Yale University
• Predictions are the holy grail of BI systems and initiatives.
• Most all corporations have need for forecasting.
• Typical forecasting systems
  • Are stand alone or from ERP (not integrated to BI system)
  • Tend to use straight line or heuristic calculations.
  • Not always integrated into the business.
  • Are often tied directly to the budgeting process
• High level of angst surrounding forecasts.
Forecasting Should...

• Should be integrated with rest of BI system.
• Should be another series of measures that are revealed in the context of historic information.
• Should be a part of the Common Enterprise Model.
• Should have visibility across functional areas and roles in corporations
• Should leverage most powerful calculation tools (database and BI system)
• Ideally adjusted based on an integrated view across corporate functions (marketing, operations, finance, etc.)
• Rule-based heuristic (last period, last period +5%, etc.)
• Cross-sectional methodologies (point in time)
• Time series (time sequenced data series)
• Mixed models
• Averages (moving, weighted, etc.)
• Linear and Non-linear regressions (line fitting)
• Transforms, projections, min/max
Methodologies for Today

• OLAP Geneva Forecasting Engine
  • Holt Winters for time series
• Oracle R Enterprise
  • ARIMA
• ODM Classification and Regression (overview)
• OBIEE Time Series Functions (overview)
• FCOPEN function -- Creates a forecasting context.
• FCSET command -- Specifies the forecast characteristics.
• FCEXEC command -- Executes a forecast and populates Oracle OLAP variables with forecasting data.
• FCQUERY function -- Retrieves information about the characteristics of a forecast or a trial of a forecast.
• FCCLOSE command -- Closes a forecasting context.
METHOD ‘method’

- **AUTOMATIC** best fit for the data. (Default)
- **LINREG** linear regression \((y=a\times x+b)\) is fitted to the data.
- **NLREG1** nonlinear regression \(x' = \log(x)\) and \(y' = \log(y)\) a polynomial model between \(x\) and \(y(y=c\times x^a)\).
- **NLREG2** nonlinear regression \(x'=x\) and \(y'=\ln(y)\) an exponential model between \(x\) and \(y(y=c\times e^{ax})\).
- **NLREG3** nonlinear regression \(x'=\log(x)\) and \(y'=y\) a logarithmic model between \(x\) and \(y(y=a\times \log(x)+b)\).
- **NLREG4** nonlinear regression method \(ix'=1/x\) and \(y'=1/y\) an asymptotic curve \((y=x/(a+bx))\).
- **NLREG5** nonlinear regression method \(x'=x\) and \(y'=\ln(y/(K-y))\) an exponential asymptotic curve \((y=cKe^{ax}/(1+ce^{ax}))\).
- **SESMOOTH** single exponential smoothing method intended for short term forecasts of non-seasonal data.
- **DESMOOTH** double exponential smoothing method exponential smoothing is applied to both the series and the trend term.
- **CROSTON** Croston's Intermittent Demand method. used for intermittent data where more than half of the observations are zero
- **HOLT/WINTERS** “triple” exponential smoothing. used on seasonal data
Using “Holt-Winters”

• Triple “Exponential Smoothing” methodology
• Used for data suspected to be seasonal
• Needs multiple seasons
• Assumes regular periods
• Pre/post processing may be necessary (fiscal calendar 445, irregular holidays, “Black Swans”, outages, etc.)
Exponential Smoothing

- Methodology for smoothing data and preferencing more recent periods when doing time series forecasts.
- Similar conceptually to a weighted moving average.
- Weights decline according to an exponential function. 
  \{1, (1-\alpha), (1-\alpha)^2, (1-\alpha)^3, \ldots\}
- Higher values give more weight to more recent periods.

- Single (weighted average of most recent observation and the most recent smoothed statistic)
- Double (trend either up or down)
- Triple (period effect)
FCSET Parameters

- **ALLOCLAST** \{YES\|NO\}
- **ALPHA** \{MAX\|MIN\|STEP\} decimal
- **APPROACH** \{'APPAUTO'\|'APPMA NUAL'\}
- **BETA** \{MAX\|MIN\|STEP\} decimal
- **COMPSMOOTH** \{YES\|NO\}
- **CYCDECFAC** \{MAX\|MIN\} decimal
- **HISTPERIODS** integer
- **MAXFACTOR** decimal
- **METHOD** 'method'
- **MINFCFACTOR** decimal
- **MPTDECAY** \{MAX\|MIN\} decimal
- **NTRIALS** integer
- **PERIODICITY** cycle-spec
- **RATIO** decimal
- **SMOOTHING** \{YES\|NO\}
- **TRANSFORM** \{'TRNOSEA'\|'TRSEA '\|'TRMPT'\}
- **TRENDHOLD** \{MAX\|MIN\| STEP\} decimal
- **WINDOWLEN** integer
Alpha, Beta, Gamma Setting

- Default Max is 0.3
- Default Min is 0.1
- Default Step is 0.1 (0.05 <= divisible value <= 0.2)
- Greater value means nearer periods have more weight.
- Lower value means periods have more equal weight.
Recommendations

• Be careful of accepting the APPAUTO setting.
• Be aware of OLAP’s 13th month
• Match HISTPERIODS with PERIODICITY for best results
• PERIODICITY cycle-spec is hierarchical from higher grain to lower
  • Ex {52,7} 52 weeks in a year, 7 days in a week
  • Ex {4,13,7} 4 quarters in a year, 13 weeks in a quarter, 7 days in a week
  • Ex {12} 12 months in a year
• Months are challenging to incorporate with other periods
Common Transformations

• Use average value per period to eliminate differences among periods (especially months)
• Shorter periods can reveal interesting patterns (e.g. average daily sales rather than average)
Essbase @TREND

• Includes single, double, and triple exponential smoothing techniques.
• Includes linear and non-linear regression option.
• Does not include an auto-choice function.
• Non-linear regression transforms must be manually applied.

• Many other transform, calculation, and modeling capabilities in Essbase.
ARIMA

- Autoregressive Integrated Moving Average
- Powerful algorithm for series analysis and prediction
- Three parameters \((p,d,q)\)
  - Auto regression (how reliant series values are on previous series values). AR(0) is white noise.
  - Integrated (degree of AR differencing, Random Walk)
  - Moving average (smoothing function)
- ARIMA \((1,0,0) = AR(1)\)
- ARIMA \((1,0,1) = ARMA (1,1)\)
- Large number of potential models
- Know the name Rob Hyndman for ARIMA in R
Stationarity

• Processes with no growth related to time.
• Random walks are stationary.
• Necessary to difference non-stationary series before applying ARMA models. (ARIMA handles this through the “Integrated” term \(d\) of the \((p,d,q)\) model parameters.)
Non-Seasonal ARIMA \((p,d,q)\)

- \(\phi(B)(1-B^d)\gamma_t = c + \theta(B)\varepsilon_t\)
- \(\{\varepsilon_t\}\) is a white noise process with 0 mean and variance \(\sigma^2\).
- \(B\) is a backshift operator
- \(\phi(z)\) is a polynomial of order \(p\)
- \(\theta(z)\) is a polynomial of order \(q\)
Seasonal ARIMA \((p,d,q)(P,D,Q)\downarrow m\)

- \(\Phi(B^m)\phi(B)(1-B^D)(1-B^d)\gamma t = c + \Theta(B^m)\Theta(B)\varepsilon t\)
- \(\{\varepsilon t\}\) is a white noise process with 0 mean and variance \(\sigma^2\).
- \(B\) is a backshift operator
- \(\Phi(z)\) is a polynomial of order \(p\)
- \(\Theta(z)\) is a polynomial of order \(q\)
Forecast() function in R

Includes methods:
- ets()
- auto.arima()
- Arima()
- arima()
- HoltWinters()
- StructTS()

Produces
- the original series;
- point forecasts;
- prediction intervals of specified coverage;
- the forecasting method used and information about the fitted model;
- residuals from the fitted model;
- one-step forecasts from the fitted model for the period of the observed data.
Choosing an ARIMA model

- Auto.arima can be used for model choice.
- Manual model choice requires hypothesis testing and evaluation of results.
- Use minimum AIC to choose the best model
  - $AIC = -2 \log(L) + 2(p+q+P+Q+k)$
  - Compare AIC values to each other, absolute values carry no meaning.
ARIMA vs. Holt-Winters

- Holt-Winters can be used for series that are seasonal and have a trend. (require order 2 differencing in ARIMA)
- Model selection can be complex in ARIMA and auto.arima selection may not be well understood.
- ARIMA best for stationary data series.
- ARIMA very powerful, but more to learn.
- Initial values more important in ARIMA (can have a big effect on predictions depending on model selected.)
- ARIMA provides confidence intervals
• Very powerful, accessible capability
• Time dimension must be designated
• Query results must be exact to pull from cache
• Can be “expensive” in processing
• Make sure that unique keys are defined at each level (“Jan13” rather than “Jan”)
AGO function

- Defines a time-based offset
- Can nest multiple AGO statements (same level)
- `Ago<<Measure>>, <<Level>>, <<Number of Periods>>`
  - Measure is a fact such as sales.
  - Level is an optional term, default is set by the grain of the query (BY clause) or is specified in repository for level based measures.
  - Number of periods is an integer specifying the offset value.
• Time-based aggregation function.
• Calculates based on starting value to current.
• Can nest with AGO (same level)
• ToDate(<Measure>, <Level>)
• Measure is a fact such as sales
• Level is the time grain such as year or month
PERIODROLLING

• Defines a period of time contextually
• Perform an operation across a specified set of query grain periods
  PeriodRolling(<<Measure>>, <<Starting Period Offset>>, <<Ending Period Offset>>, <<[Hierarchy]>>)
• Measure is a fact such as sales
• Starting Period Offset is an integer value, use a minus sign (“-2” means 2 periods ago)
• Ending Period Offset defines the end of the period, use a zero for current period
• Hierarchy is an optional setting to specify which time hierarchy to use such as “fiscal”
• Use “unbound” for starting period offset to calculate total from beginning
• PeriodRolling uses either the query level grain of “measure” or the measure level for “measure” if it has been set in the Admin tool.
Oracle Data Mining

• Oracle Data Mining is an option for the Enterprise Edition of the Oracle Database.
• A collection of APIs and specialized SQL functions.
• Includes a large number of specialized algorithms and built-in procedures.
• Makes use of many built-in capabilities of the Oracle Database.
• ODM typically refers to “Oracle Data Mining”
## Oracle Data Mining Algorithms

<table>
<thead>
<tr>
<th>Problem</th>
<th>Algorithm</th>
<th>Applicability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>Logistic Regression (GLM)</td>
<td>Classical statistical technique</td>
</tr>
<tr>
<td></td>
<td>Decision Trees</td>
<td>Popular / Rules / transparency</td>
</tr>
<tr>
<td></td>
<td>Naïve Bayes</td>
<td>Embedded app</td>
</tr>
<tr>
<td></td>
<td>Support Vector Machine</td>
<td>Wide / narrow data / text</td>
</tr>
<tr>
<td>Regression</td>
<td>Multiple Regression (GLM)</td>
<td>Classical statistical technique</td>
</tr>
<tr>
<td></td>
<td>Support Vector Machine</td>
<td>Wide / narrow data / text</td>
</tr>
<tr>
<td>Anomaly Detection</td>
<td>One Class SVM</td>
<td>Fraud Detection</td>
</tr>
<tr>
<td>Attribute</td>
<td>Minimum Description Length (MDL)</td>
<td>Attribute reduction</td>
</tr>
<tr>
<td>Importance</td>
<td></td>
<td>Identify useful data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reduce data noise</td>
</tr>
<tr>
<td>Association</td>
<td>Apriori</td>
<td>Market basket analysis</td>
</tr>
<tr>
<td>Rules</td>
<td></td>
<td>Link analysis</td>
</tr>
<tr>
<td>Clustering</td>
<td>Hierarchical K-Means</td>
<td>Product grouping</td>
</tr>
<tr>
<td></td>
<td>Hierarchical O-Cluster</td>
<td>Text mining</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gene and protein analysis</td>
</tr>
<tr>
<td>Feature Extraction</td>
<td>NMF</td>
<td>Text analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Feature reduction</td>
</tr>
</tbody>
</table>
Classification

• Prediction model for non-continuous information
  • Binary such as yes/no
  • Limited set (low/medium/high)

• Involves “supervised learning”
  • Prediction directed by a previously known dependent variable or “target” variable.
  • Commonly includes three phases:
    • Training
    • Testing
    • Scoring

• Results in predictive models that are applied to new data sets.

• In our example, we predict which prospects are likely to buy insurance.
Oracle Test Drive

- Free to try out Oracle BI
- Go to www.vlamis.com/testdrive-registration/
- Runs off of Amazon AWS
- Hands-on Labs based on Collaborate 2012 HOLs
- Test Drives for:
  - Oracle BI
  - BI Publisher
  - Microsoft Excel against Oracle OLAP
  - Oracle Data Mining
  - Map Views in OBIEE
- Once sign up, you have private instance for 5 hours
- Available now