Implementing a Predictive Model

Danny Wijnschenk
Sales Engineer

CacheConf
Data Analytics

Descriptive Analytics

What?

now

Predictive Analytics

What?

Prescriptive Analytics

Why?
# Examples of Predictive Analytics

## Weather

### Moscow, Russia Weather

<table>
<thead>
<tr>
<th>DAY</th>
<th>COND</th>
<th>HIGH</th>
<th>LOW</th>
<th>DESCRIPTION</th>
<th>PRECIP</th>
<th>WIND</th>
</tr>
</thead>
<tbody>
<tr>
<td>TODAY</td>
<td></td>
<td>13°C</td>
<td>7°</td>
<td>Showers/Wind</td>
<td>50%</td>
<td>WNW 34 km/h</td>
</tr>
<tr>
<td>May 18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TUE</td>
<td></td>
<td>13°</td>
<td>8°</td>
<td>Cloudy</td>
<td>20%</td>
<td>W 26 km/h</td>
</tr>
<tr>
<td>May 19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WED</td>
<td></td>
<td>23°</td>
<td>11°</td>
<td>Partly Cloudy</td>
<td>0%</td>
<td>NW 13 km/h</td>
</tr>
<tr>
<td>May 20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>THU</td>
<td></td>
<td>27°</td>
<td>15°</td>
<td>Partly Cloudy</td>
<td>20%</td>
<td>S 16 km/h</td>
</tr>
<tr>
<td>May 21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FRI</td>
<td></td>
<td>27°</td>
<td>15°</td>
<td>Partly Cloudy</td>
<td>20%</td>
<td>WSW 21 km/h</td>
</tr>
<tr>
<td>May 22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Examples of Predictive Analytics

Retail

Amazon Knows What You Want Before You Buy It

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Walmart Labs Buys Data Analytics And Predictive Intelligence Startup Inkuru

Posted Jun 10, 2013 by Leena Rao (@leenarao)
Examples of Predictive Analytics

Energy Smart Meters
Building Predictive Models

1. Explore
2. Clean
3. Train
4. Test
5. Deploy

InterSystems logo is visible at the top left corner of the page.
Building Predictive Models

**Explore**

- Get the historical data
- External data needed?
- Look for dependencies (Overfitting)
Building Predictive Models

Clean

- Empty fields
- Wrong data
- Formats
Building Predictive Models

Train

- Divide data in training and testing set
- Choose model
- Train the model with training set
Building Predictive Models

- Use test set to verify results
- Go back to Explore-Clean-Train till metrics are OK
Data Scientists

- precision
- probability
- statistics
- recall
- outliers

- R
- SAS
- SPSS
- KNIME
Production Environment
PMML

- Predictive Model Markup Language
- XML based file format
- Supported by major modelling tools (SPSS, SAS, R, Knime, MicroStrategy, ...)
- Connects the Data Scientists World to the Production World
Deploy a Predictive Model in Caché

- Insert PMML in a class or use Wizard
- Compile
- Run
  - In Caché Object Script
  - In DeepSee (as Dimension, Calculated Measure, KPI)
  - In Caché SQL
Demo
Steps of the Demo

• Show how to build a Predictive Model
  – Using Knime (www.knime.org)
    • Open Source
    • User Friendly UI

• Export the model in PMML

• Import PMML in Caché

• Use the model in Caché (COS, SQL, DeepSee)
Case 1: Predicting the Origin of Cars

• Create model based on Decision Tree
• Car data contains
  – Brand, Model, Year of car
  – Number of Cylinders of engine
  – Weight of the car
  – Engine Horsepower
  – Displacement of Cylinders
  – Consumption in Miles per Gallon
• Origin should be Japan, US or Europe
Step 1: Database Reader
Step 2: Column Filter
Step 3: Decision Tree Learner

Database Reader → Column Filter → Decision Tree Learner → PMML Writer

Dialog - 0:3 - Decision Tree Learner

- **Class column**: Origin
- **Quality measure**: Gini index
- **Pruning method**: No pruning
- **Reduced Error Pruning**: Check box selected
- **Min number records per node**: 2
- **Number records to store for view**: 10,000
- **Average split point**: Check box selected
- **Number threads**: 8
- **Skip nominal columns without domain information**: Check box selected
Step 4: View Result
Step 4: Example of Overfitting
Step 5: Export PMML

```
<MinningField name="Horsepower" invalidValueTreatment="asIs"/>
<MinningField name="MilesPerGallon" invalidValueTreatment="asIs"/>
<MinningField name="ModelYear" invalidValueTreatment="asIs"/>
<MinningField name="Weight" invalidValueTreatment="asIs"/>
<MinningField name="Origin" invalidValueTreatment="asIs" usageType="predicted"/>
</MinningSchema>
<Node id="0" score="US" recordCount="266.0">
</Node id="1" score="Asia" recordCount="115.0">
</Node id="2" score="Asia" recordCount="107.0">
</Node id="3" score="Asia" recordCount="53.0">
```
Case 2 : Predicting Car Consumption

• Create model based on Polynomial Regression

• Car data contains
  – Brand, Model, Year, Origin of car
  – Number of Cylinders of engine
  – Weight of the car
  – Engine Horsepower
  – Displacement of Cylinders
  – Consumption in Miles per Gallon
**Step 1: Database Reader**

[Image of a KNIME workflow diagram showing the steps and a table with data columns such as Acceleration, Brand, Cylinders, Displacement, Horsepower, Miles Per Gallon, Model, and Origin.]
Step 2 : Cleanup the Data
Step 3: Polynomial Regression Learner
Step 4: View Result
Step 5 : Export PMML

```
<MiningField name="ModelYear" invalidValueTreatment="asIs"/>
<MiningField name="Weight" invalidValueTreatment="asIs"/>
<MiningField name="MilesPerGallon" invalidValueTreatment="asIs" usageType="predicted"/>
</MiningSchema>
<RegressionTable intercept="468.8287211806919">
  <NumericPredictor name="Acceleration" coefficient="-1.8277555671414802"/>
  <NumericPredictor name="Acceleration" exponent="2" coefficient="0.05151120682898025"/>
  <NumericPredictor name="Cylinders" coefficient="4.614001180239825"/>
  <NumericPredictor name="Cylinders" exponent="2" coefficient="-0.29860315770274326"/>
  <NumericPredictor name="Displacement" coefficient="-0.07795045961131564"/>
  <NumericPredictor name="Displacement" exponent="2" coefficient="1.198142310369632E-4"/>
  <NumericPredictor name="Horsepower" coefficient="-0.10646731518352937"/>
  <NumericPredictor name="Horsepower" exponent="2" coefficient="1.574027060684824E-4"/>
  <NumericPredictor name="ModelYear" coefficient="-11.437811396680956"/>
  <NumericPredictor name="ModelYear" exponent="2" coefficient="0.08066603594487465"/>
  <NumericPredictor name="Weight" coefficient="-0.01309352621142482"/>
  <NumericPredictor name="Weight" exponent="2" coefficient="1.5166483916730523E-6"/>
</RegressionTable>
</RegressionModel>
<PMMML>
```
Import PMML in Caché Class

```caché
Class PredictiveAnalytics.ModelOrigin Extends %DeepSee.PMML.Definition
{

  XData PMML
  {
    <?xml version="1.0" encoding="UTF-8"?>
    <PMML version="4.1" xmlns="http://www.dmg.org/PMML-4_1">
      <Header copyright="dwijnsch">
        <Application name="KNIME" version="2.9.0"/>
      </Header>
      <DataDictionary numberOfFields="10">
        <DataField name="_ID" optype="continuous" dataType="integer">
          <Interval closure="closedClosed" leftMargin="1.0" rightMargin=""/>
        </DataField>
        <DataField name="Acceleration" optype="continuous" dataType="double">
          <Interval closure="closedClosed" leftMargin="8.5" rightMargin=""/>
        </DataField>
        <DataField name="Cylinders" optype="continuous" dataType="integer">
          <Interval closure="closedClosed" leftMargin="3.0" rightMargin=""/>
        </DataField>
      </DataDictionary>
    </PMML>
  }
}
```
Run model in Caché Object Script

Do ##class(PredictiveAnalytics.ModelOrigin).%GetModelInstance(, .model)

Set data("Acceleration") = 12
Set data("MilesPerGallon") = 20
Set data("Weight") = 2400
Set data("Horsepower") = 146

Do model.%ExecuteModel(.data, .output)

Write output.Origin
Run model in Caché SQL

```sql
SELECT PredictiveAnalytics_ModelOrigin.DecisionTree(12,,120,146,20,,3000)
```

Row count: 1 Performance: 0.000 seconds 6 global references Cached Query: %sqlcq.PREDANA.cls19 Last update: 2014-03-05 08:38:01.437

1 row(s) affected
Run model from any UI (e.g. AngularJS)

```html
<form name="frmCar" ng-controller="CarController as car" ng-submit="" class="form-horizontal">
  <fieldset>
    <div class="well">
      <label>Car Origin Prediction = {{car.Prediction}}</label>
      <div ng-repeat="trace in car.Trace">
        <label>-- {{trace.Field}} : {{trace.Msg}}
        </label>
      </div>
    </div>
  </fieldset>
</form>

<!-- Displacement -->

```html
<div class="form-group">
  <label for="inputDisplacement" class="control-label col-xs-2">Displacement</label>
  <div class="col-xs-10">
    <input type="text" class="form-control" id="inputDisplacement" placeholder="60-500"
           ng-model="car.Displacement" ng-change="predictOrigin(car)"
    ></div>
</div>

```javascript
app.factory('restPredictOrigin', function ($resource) {
    return $resource('http://'+document.domain+':57775/csp/predana/rest/PredictOrigin', {}, {
        update: {method: 'POST'}
    });
});
```

```javascript
{
    "Prediction": "Asia",
    "Trace": [
        {
            "Field": "Displacement",
            "Msg": "Tree node 1: (Displacement <= 134.5)"
        },
        {
            "Field": "Weight",
            "Msg": "Tree node 2: (Weight <= 2757.5)"
        },
        {
            "Field": "Displacement",
            "Msg": "Tree node 3: (Displacement <= 97.25)"
        }
    ]
}
```
Test model in DeepSee

Test results

Record count: 132

Precision: 85.77% (75.45)

Recall: 83.33% (76.67)

F-measure: 84.53% (76.06)

Details by actual value

Overview

ASIA

EUROPE

US

Details by actual value

Overview

Details

Actual value: EUROPE (26)

P: 78%  R: 54%  F: 64%

true positive: 14  false negative: 12

false positive: 4  true negative: 102

Item: EUROPE  Value: 14

Series count

US (12%)

ASIA (35%)
Run model in DeepSee

<table>
<thead>
<tr>
<th>Origin</th>
<th>Asia</th>
<th>Europe</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asia</td>
<td>73</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Europe</td>
<td>11</td>
<td>55</td>
<td>4</td>
</tr>
<tr>
<td>US</td>
<td>6</td>
<td>4</td>
<td>239</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#</th>
<th>Brand</th>
<th>Model</th>
<th>ModelYear</th>
<th>Cylinders</th>
<th>Acceleration</th>
<th>Displacement</th>
<th>Horsepower</th>
<th>MilesPerGallon</th>
<th>Weight</th>
<th>Origin</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>chevrolet</td>
<td>chevelle malibu</td>
<td>70</td>
<td>8</td>
<td>12</td>
<td>307</td>
<td>130</td>
<td>18</td>
<td>3504</td>
<td>US</td>
<td>15.55366342340926946</td>
</tr>
<tr>
<td>2</td>
<td>buick</td>
<td>skylark 320</td>
<td>70</td>
<td>8</td>
<td>11.5</td>
<td>350</td>
<td>165</td>
<td>15</td>
<td>3693</td>
<td>US</td>
<td>13.48240835956640139</td>
</tr>
<tr>
<td>3</td>
<td>plymouth</td>
<td>satellite</td>
<td>70</td>
<td>8</td>
<td>11</td>
<td>318</td>
<td>150</td>
<td>18</td>
<td>3436</td>
<td>US</td>
<td>15.18966953569111001</td>
</tr>
<tr>
<td>4</td>
<td>amc</td>
<td>rebel sst</td>
<td>70</td>
<td>8</td>
<td>12</td>
<td>304</td>
<td>150</td>
<td>16</td>
<td>3433</td>
<td>US</td>
<td>14.50266322076093966</td>
</tr>
<tr>
<td>5</td>
<td>ford</td>
<td>torino</td>
<td>70</td>
<td>8</td>
<td>10.5</td>
<td>302</td>
<td>140</td>
<td>17</td>
<td>3449</td>
<td>US</td>
<td>16.18218746220322668</td>
</tr>
<tr>
<td>6</td>
<td>ford</td>
<td>galaxie 500</td>
<td>70</td>
<td>8</td>
<td>10</td>
<td>429</td>
<td>198</td>
<td>15</td>
<td>4341</td>
<td>US</td>
<td>13.56134534923182278</td>
</tr>
<tr>
<td>7</td>
<td>chevrolet</td>
<td>impala</td>
<td>70</td>
<td>8</td>
<td>9</td>
<td>454</td>
<td>220</td>
<td>14</td>
<td>4354</td>
<td>US</td>
<td>14.21293686432677949</td>
</tr>
<tr>
<td>8</td>
<td>plymouth</td>
<td>fury iii</td>
<td>70</td>
<td>8</td>
<td>8.5</td>
<td>440</td>
<td>215</td>
<td>14</td>
<td>4312</td>
<td>US</td>
<td>14.45570465893638744</td>
</tr>
<tr>
<td>9</td>
<td>pontiac</td>
<td>catalina</td>
<td>70</td>
<td>8</td>
<td>10</td>
<td>455</td>
<td>225</td>
<td>14</td>
<td>4425</td>
<td>US</td>
<td>13.22843060468615661</td>
</tr>
<tr>
<td>10</td>
<td>amc</td>
<td>ambassador dpl</td>
<td>70</td>
<td>8</td>
<td>8.5</td>
<td>390</td>
<td>190</td>
<td>15</td>
<td>3850</td>
<td>US</td>
<td>14.77914055523089386</td>
</tr>
</tbody>
</table>
Run model in DeepSee
TakeAway Summary

• Data Scientists make Predictive Models with specialized tools
• Predictive Models can be exported in a PMML description
• Caché can deploy a Predictive Model by using the PMML description
• An Application Developer can run the model anywhere in Caché (COS, SQL, DeepSee)
Questions?