Oracle Advanced Analytics for Fraud and Anomaly Detection

Make Big Data + Analytics Simple

Charlie Berger, MS Engineering, MBA
Sr. Director Product Management, Data Mining and Advanced Analytics
charlie.berger@oracle.com www.twitter.com/CharlieDataMine

ORACLE

Safe Harbor Statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

Fraud Statistics
By The Numbers

Coalition Against Insurance Fraud

- Overall
  - Conservatively, **fraud steals $80 billion a year across all lines of insurance**. (Coalition Against Insurance Fraud est.).
  - Fraud comprises about 10 percent of property-casualty insurance losses and loss adjustment expenses each year.
- Fraud costs for insurers
  - **Fraud accounts for 5-10 percent of claims costs** for U.S. and Canadian insurers. Nearly one-third of insurers (32 percent) say fraud was as high as 20 percent of claims costs;
  - About 35 percent say fraud costs their companies 5-10 percent of claim volume. **More than 30 percent say fraud losses cost 10-20 percent of claim volume**;
  - Detecting fraud before claims are paid, and **upgrading analytics**, were mentioned most often as the insurers' main fraud-fighting priorities;

http://www.insurancefraud.org/statistics.htm#.VfreKRFVhBc

Fraud Statistics
By The Numbers

Coalition Against Insurance Fraud

- Medicare & Medicaid
  - Nearly **$80 billion of improper Medicare and Medicaid payments** were made in FY 2014;
  - Anti-fraud efforts **recovered $3.3 billion in taxpayer dollars** in FY 2014; and
  - $7.70 was returned for every anti-fraud dollar invested. This is about $2 higher than the average ROI since 1997. It's also the third-highest ROI. (U.S. Department of HHS, March 2015)
- Automobile - Bodily injury claims
  - Staged-crash rings **fleece auto insurers out of billions of dollars a year by billing for unneeded treatment of phantom injuries**. Usually these are bogus soft-tissue injuries such as sore backs or whiplash, which are difficult to medically identify and dispute.
- Hotspot states
  - **Drivers in Lawrence , MA— the "worst hotbed of fraudulent claims"** — have saved more than $68 million; Larger chiropractors in Lawrence have decreased in both clinic counts and billings by up to 90 percent. High-volume physical therapy clinics (billings exceeding $100,000 annually) have been eliminated, and attorney involvement in PIP claims has dropped;

ORACLE

http://www.insurancefraud.org/statistics.htm#.VfreKRFVhBc

People's Attitudes About Fraud

Consumers

Coalition Against Insurance Fraud

- **Nearly one of four Americans say it's ok to defraud insurers**
  - Some **8 percent say it's "quite acceptable" to bilk insurers**, while 16 percent say it's "somewhat acceptable."
  - About **one in 10 people agree** it's ok to submit claims for items that aren't lost or damaged, or for personal injuries that didn't occur.
  - **Two of five people are "not very likely" or "not likely at all" to report someone who ripped of an insurer**. Accenture Ltd.(2003)
- **Nearly one of 10 Americans would commit insurance fraud** if they knew they could get away with it.
- Nearly three of 10 Americans (29 percent) wouldn't report insurance scams committed by someone they know. Progressive Insurance (2001)

ORACLE

American Society of Certified Fraud Examiners

20 Ways to Detect Fraud

1. Unusual Behavior

The perpetrator will often display unusual behavior, that when taken as a whole is a strong indicator of fraud. The fraudster may not ever take a vacation or call in sick in fear of being caught. He or she may not assign out work even when overloaded. Other symptoms may be changes in behavior such as increased drinking, smoking, defensiveness, and unusual irritability and suspiciousness.

2. Complaints

Frequently tips or complaints will be received which indicate that a fraudulent action is going on. Complaints have been known to be some of the best sources of fraud and should be taken seriously. Although all too often, the motive behind the complaint may be suspect, the allegations usually have merit that warrant further investigation.

3. Stale Items in Reconciliations

In bank reconciliations, deposits or checks included in the reconciliation could be indicative of theft. Missing deposits could mean the perpetrator absconded with the money. Missing checks could indicate one made out to a bogus payee.

4. Excessive Voids

Voided sales slips could mean that the sale was rung up, the payment diverted to the use of the perpetrator, and the sales slip subsequently voided to cover the theft.

5. Missing Documents

Documents which are unable to be located can be a red flag for fraud. Although it is expected that some documents will be misplaced, the auditor should look for explanations as to why the documents are missing, and what steps were taken to locate the requested items. All too often, the auditors will select an alternate item or allow the auditee to select an alternate without determining whether or not problem exists.

6. Excessive Credit Memos

Similar to excessive voids, this technique can be used to cover the theft of cash. A credit memo to a phony customer is written out, and the cash is taken to make total cash balance.

Pretty Easy? Huh?

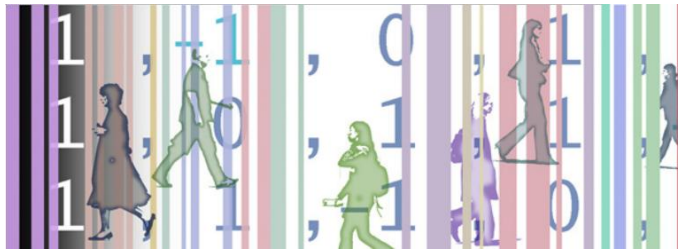
ORACLE

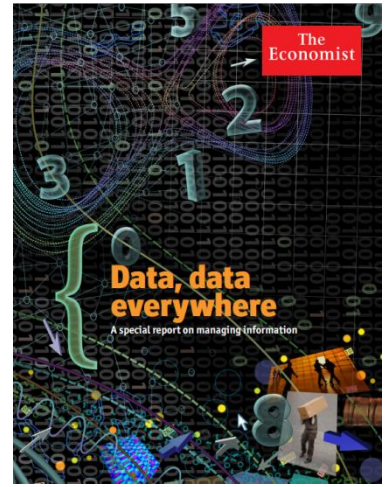
http://www.fraudiscovery.com/detect.html

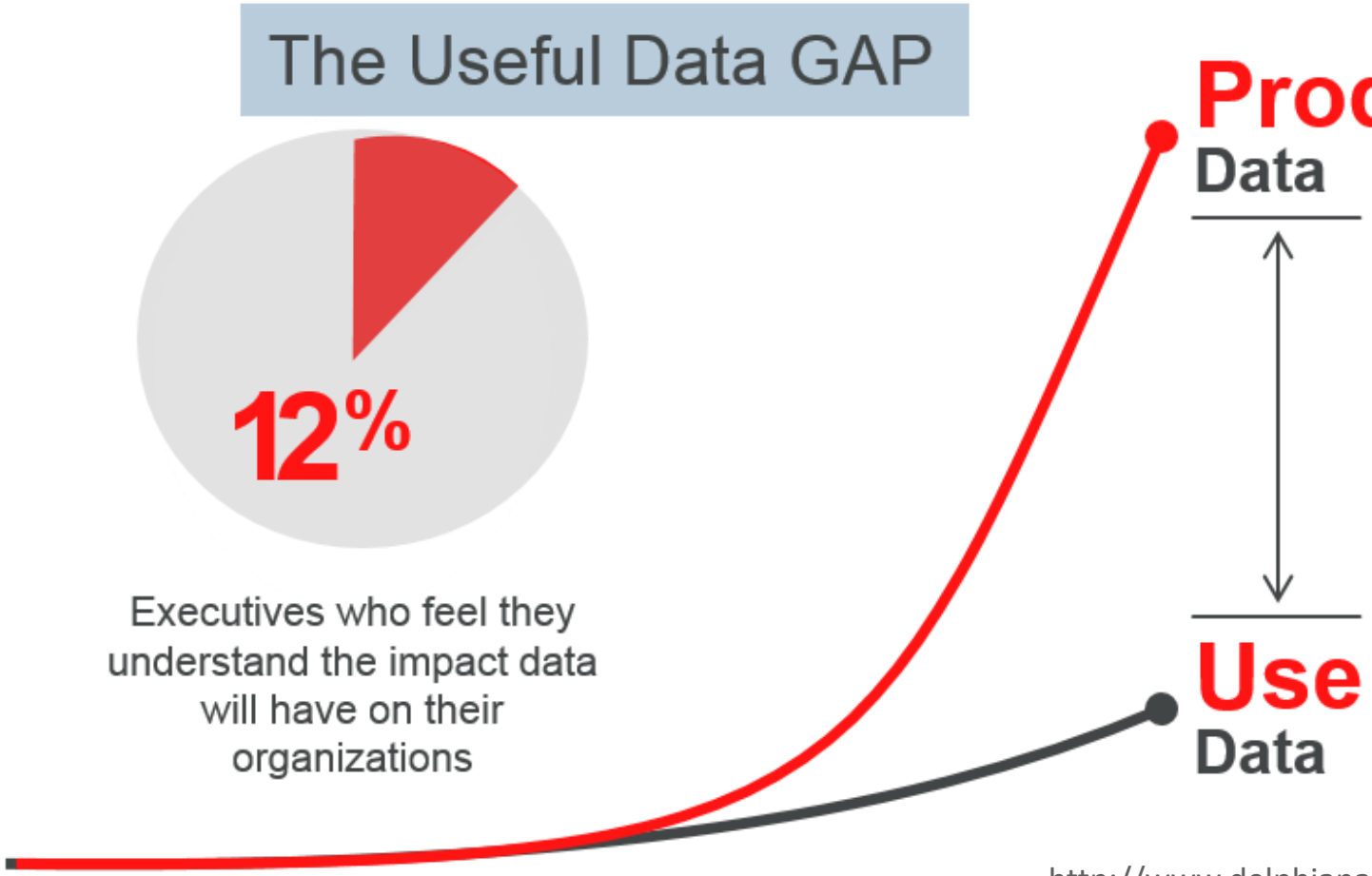
http://www.auditnet.org/testing_20_ways.htm

Data, data everywhere

Growth of Data Exponentially Greater than Growth of Data Analysts!







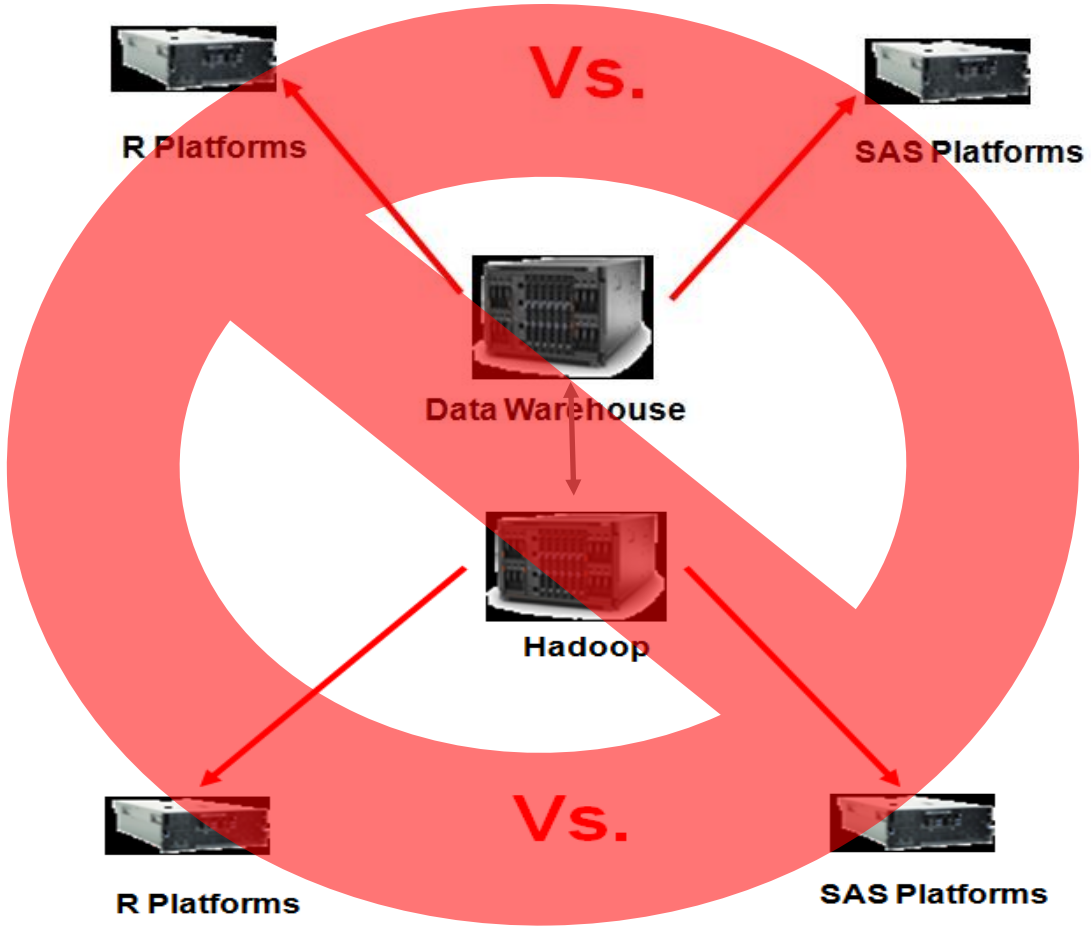
Data Analysis platforms requirements:

- Be extremely powerful and handle large data volumes
- Be easy to learn
- Be highly automated & enable deployment

http://www.delphianalytics.net/more-data-than-analysts-the-real-big-data-problem/
http://uk.emc.com/collateral/analyst-reports/ar-the-economist-data-data-everywhere.pdf

Analytics + Data Warehouse + Hadoop

- Platform Sprawl
  - More Duplicated Data
  - More Data Movement Latency
  - More Security challenges
  - More Duplicated Storage
  - More Duplicated Backups
  - More Duplicated Systems
  - More Space and Power



Vision

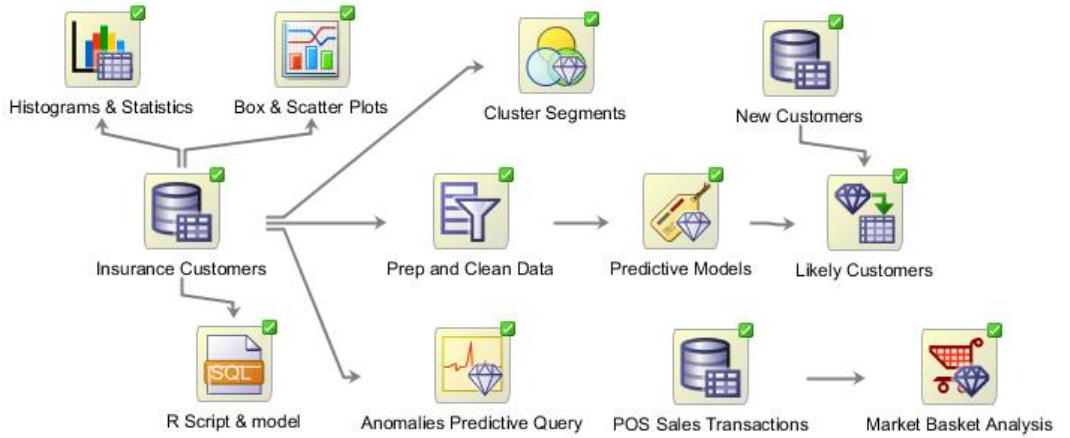
- Big Data + Analytic Platform for the Era of Big Data and Cloud
  - **Make Big Data + Analytics Model Discovery *Simple***
    - Any data size, on any computer infrastructure
    - Any variety of data (structured, unstructured, transactional, geospatial), in any combination
  - **Make Big Data + Analytics Model Deployment *Simple***
    - As a service, as a platform, as an application

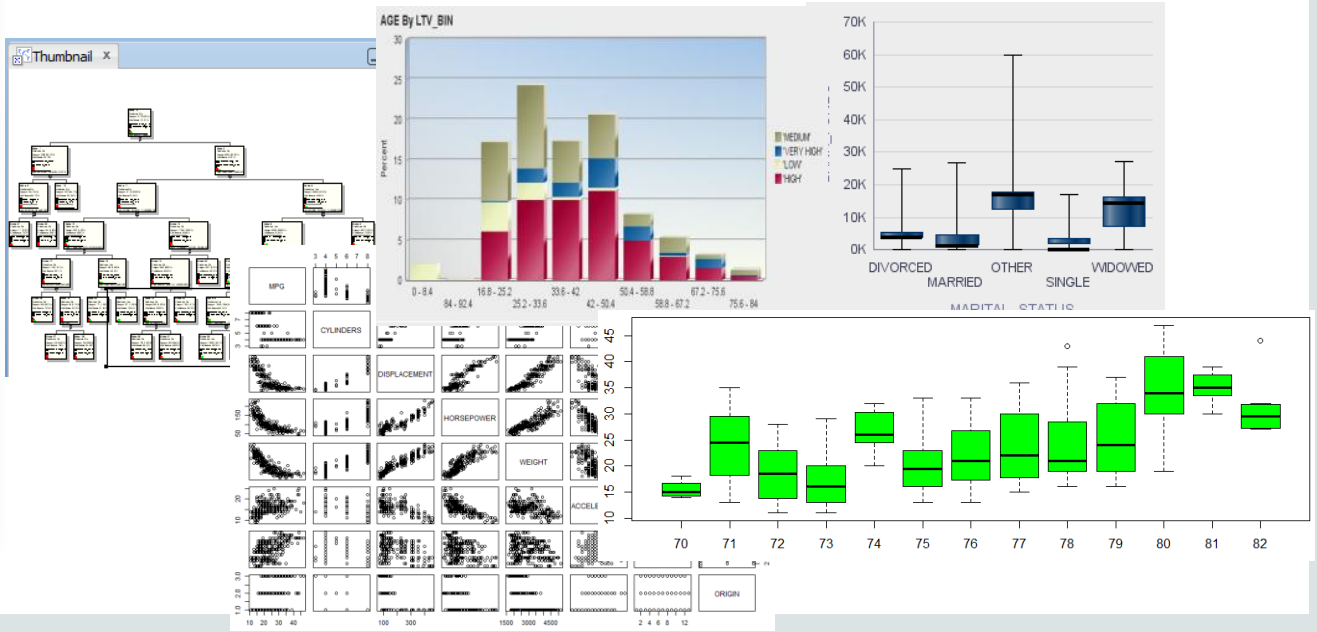
Oracle's Advanced Analytics

Fastest Way to Deliver Scalable Enterprise-wide Predictive Analytics

Key Features

- Scalable in-Database + Hadoop data mining algorithms and R integration
- Powerful predictive analytics and deployment platform
- Drag and drop workflow, R and SQL APIs
- Data analysts, data scientists & developers
- Enables enterprise predictive analytics applications



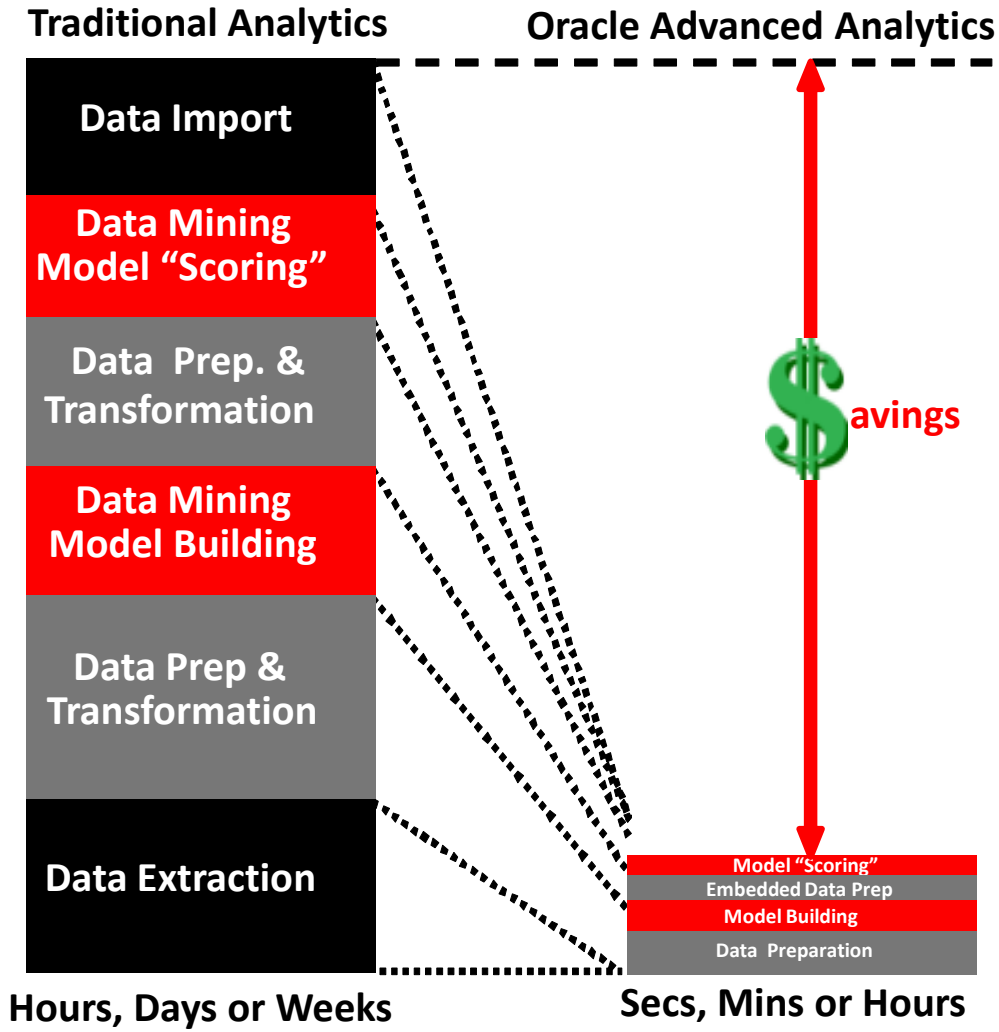


Oracle's Advanced Analytics

Fastest Way to Deliver Scalable Enterprise-wide Predictive Analytics

Major Benefits

- Data remains in Database & Hadoop
  - Model building and scoring occur in-database
  - Use R packages with data-parallel invocations
- Leverage investment in Oracle IT
  - Eliminate data duplication
  - Eliminate separate analytical servers
- Deliver enterprise-wide applications
  - GUI for Predictive Analytics & code gen
  - R interface leverages database as HPC engine



Fiserv

Risk Analytics in Electronic Payments

Objectives

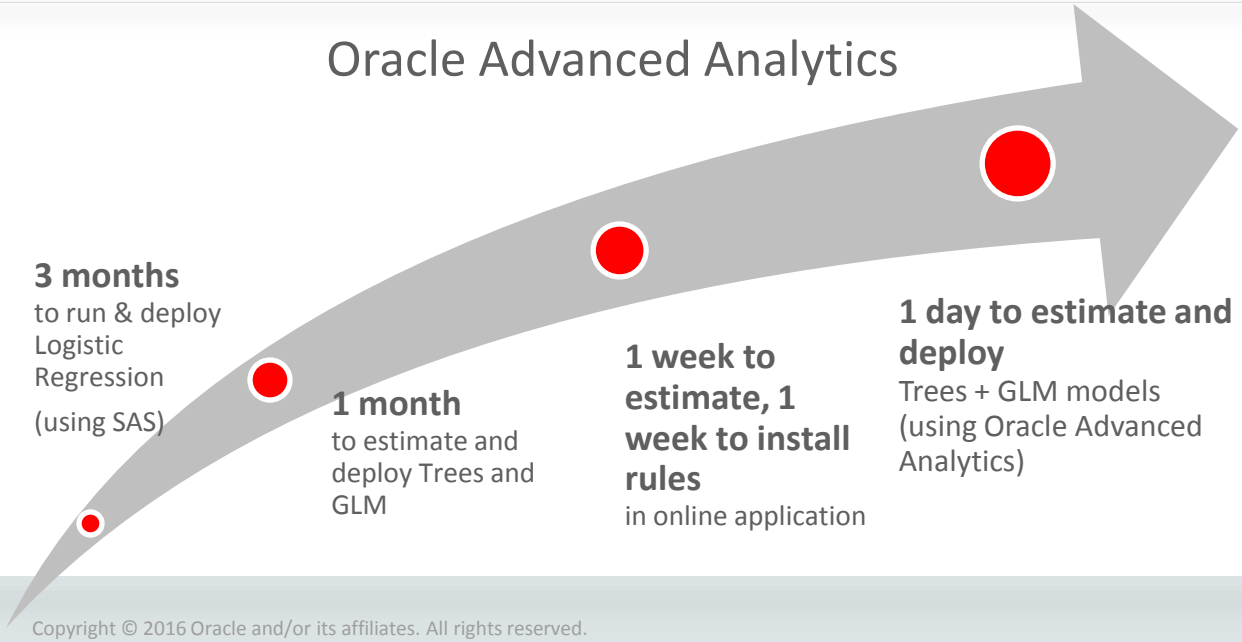
- Prevent $200M in losses every year using data to monitor, understand and anticipate fraud

Solution

- We installed OAA analytics for model development during 2014
- When choosing the tools for fraud management, speed is a critical factor
- OAA provided a fast and flexible solution for model building, visualization and integration with production processes

- "When choosing the tools for fraud management, speed is a critical factor. Oracle Advance Analytics provided a fast and flexible solution for model building, visualization and integration with production processes."

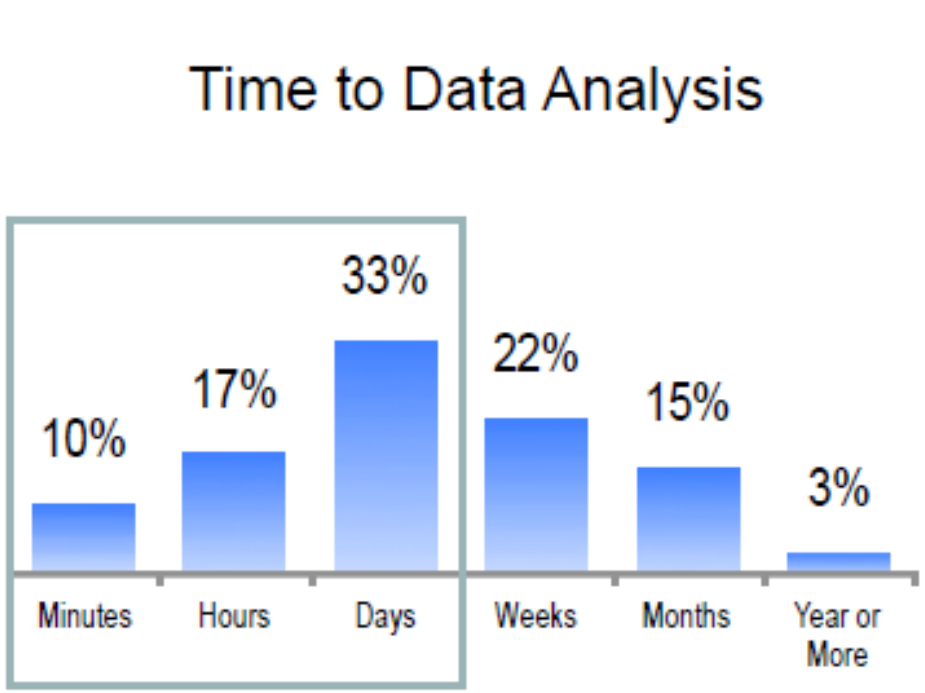
  – Miguel Barrera, Director of Risk Analytics, Fiserv Inc.
  – Julia Minkowski, Risk Analytics Manager, Fiserv Inc.

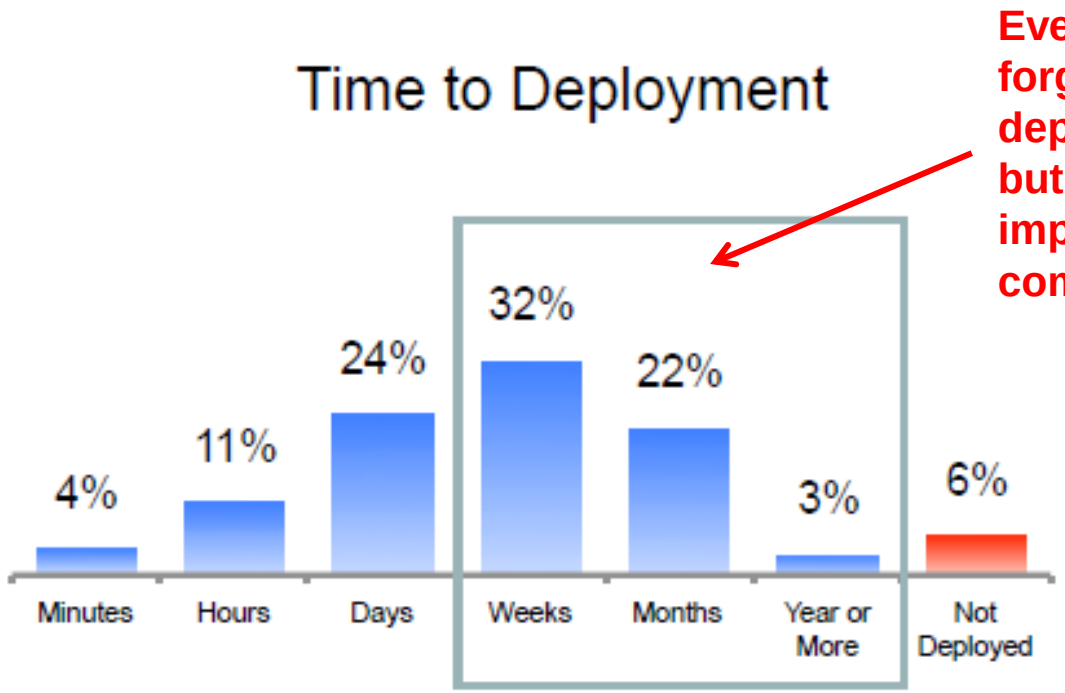


Ease of Deployment

Data Miner Survey 2016 by Rexer Analytics

While 6 out 10 data miners report the data is available for analysis within days of capture, the time to deploy the models takes substantially longer. For 60% of the respondents the deployment time will range between 3 weeks and 1year.





Everyone forgets about deployment – but is most important component!

UK National Health Service

Combating Healthcare Fraud

NHS Business Services Authority

Objectives

- Use new insight to help identify cost savings and meet goals
- Identify and prevent healthcare fraud and benefit eligibility errors to save costs
- Leverage existing data to transform business and productivity

Solution

- Identified up to GBP100 million (US$156 million) potentially saved through benefit fraud and error reduction
- Used anomaly detection to uncover fraudulent activity where some dentists split a single course of treatment into multiple parts and presented claims for multiple treatments
- Analyzed billions of records at one time to measure longer-term patient journeys and to analyze drug prescribing patterns to improve patient care

- "Oracle Advanced Analytics' data mining capabilities and Oracle Exalytics' performance really impressed us. The overall solution is very fast, and our investment very quickly provided value. We can now do so much more with our data, resulting in significant savings for the NHS as a whole"
  – Nina Monckton, Head of Information Services, NHS Business Services Authority

Oracle Exadata Database Machine

Oracle Advanced Analytics

Oracle Exalytics In-Memory Machine
Oracle Endeca Information Discovery
Oracle Business Intelligence EE

ORACLE

Oracle's Advanced Analytics

Multiple interfaces across platforms — SQL, R, GUI, Dashboards, Apps

Users

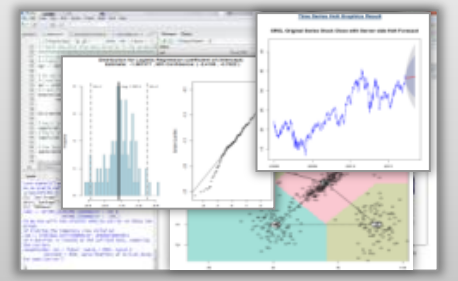
R programmers

Data & Business Analysts

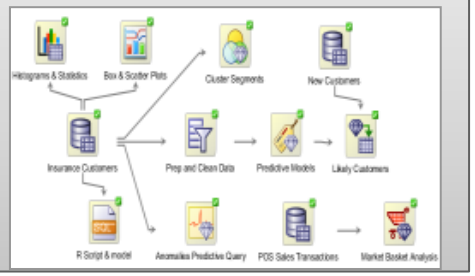
Business Analysts/Mgrs

Domain End Users

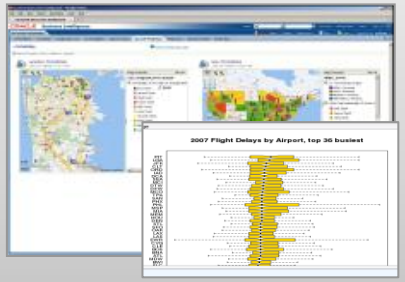
R Client



SQL Developer/ Oracle Data Miner

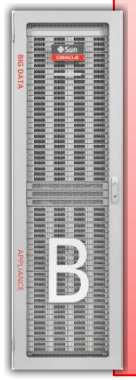


OBIEE



Applications

Platform



Hadoop

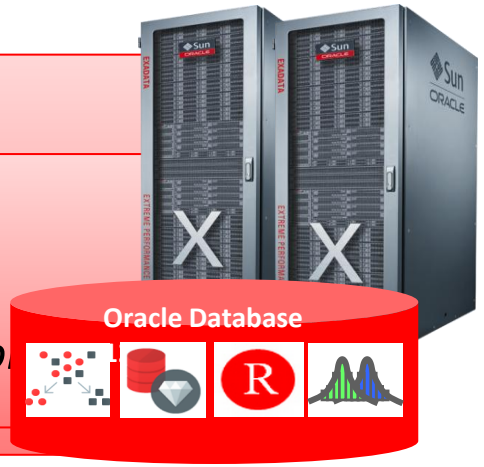
ORAAH

Parallel, distributed algorithms

Oracle Database Enterprise Edition

Oracle Advanced Analytics - Database Option

SQL Data Mining & Analytic Functions + R Integration for Scalable, Distributed, Parallel in-Database ML Executio



Oracle Cloud

Oracle Advanced Analytics Database Evolution

1998

- *7 Data Mining "Partners"*

1999

- *Oracle acquires Thinking Machine Corp's dev. team + "Darwin" data mining software*

2002

- *Oracle Data Mining 9.2i launched – 2 algorithms (NB and AR) via Java API*

2004

- *Oracle Data Mining 10gR2 SQL - 7 new SQL dm algorithms and new Oracle Data Miner "Classic" wizards driven GUI*
- *SQL statistical functions introduced*

2005

- *ODM 11g & 11gR2 adds AutoDataPrep (ADP), text mining, perf. improvements*
- *SQLDEV/Oracle Data Miner 3.2 "work flow" GUI launched*
- *Integration with "R" and introduction/addition of Oracle R Enterprise*
- *Product renamed "Oracle Advanced Analytics (ODM + ORE)*

2008

2011

- *New algorithms (EM, PCA, SVD)*
- *Predictive Queries*
- *SQLDEV/Oracle Data Miner 4.0 SQL script generation and SQL Query node (R integration)*
- *OAA/ORE 1.3 + 1.4 adds NN, Stepwise, scalable R algorithms*
- *Oracle Adv. Analytics for Hadoop Connector launched with scalable BDA algorithms*

2014

You Can Think of Oracle's Advanced Analytics Like This…

Traditional SQL

- "Human-driven" queries
- Domain expertise
- Any "*rules*" must be defined and managed

SQL Queries

- `SELECT`
- `DISTINCT`
- `AGGREGATE`
- `WHERE`
- `AND OR`
- `GROUP BY`
- `ORDER BY`
- `RANK`

+

Oracle Advanced Analytics - SQL & R

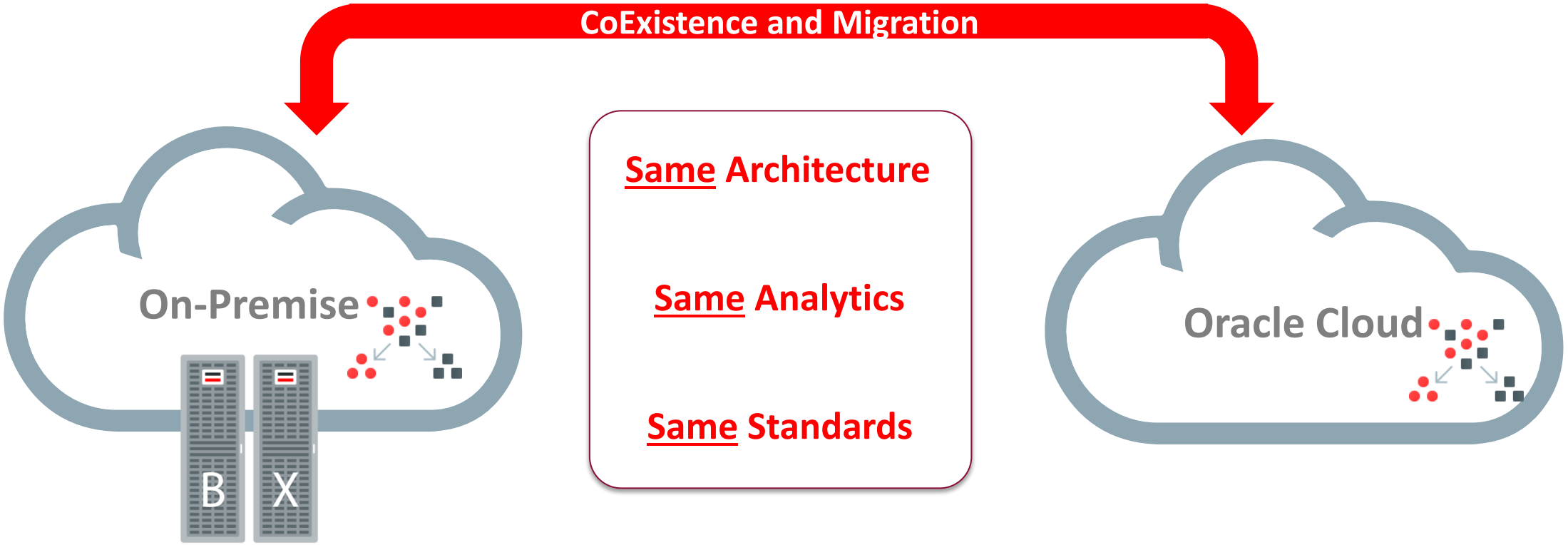
- Automated knowledge discovery, model building and deployment
- Domain expertise to assemble the "*right*" data to mine/analyze

Analytical SQL "Verbs"

- `PREDICT`
- `DETECT`
- `CLUSTER`
- `CLASSIFY`
- `REGRESS`
- `PROFILE`
- `IDENTIFY FACTORS`
- `ASSOCIATE`

Oracle Advanced Analytics—On Premise or Cloud

100% Compatibility Enables Easy Coexistence and Migration



Transparently move workloads and analytical methodologies between On-premise and public cloud

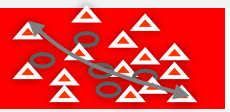
Oracle's Advanced Analytics

In-Database Data Mining Algorithms*—SQL & R & GUI Access

Classification

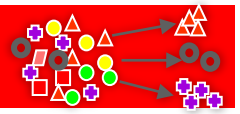
- Decision Tree
- Logistic Regression (GLM)
- Naïve Bayes
- Support Vector Machine (SVM)
- Random Forest

Regression



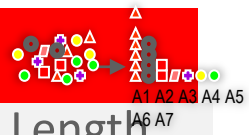
- Multiple Regression (GLM)
- Support Vector Machine (SVM)
- Linear Model
- Generalized Linear Model
- Multi-Layer Neural Networks
- Stepwise Linear Regression

Clustering



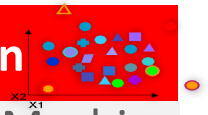
- Hierarchical k-Means
- Orthogonal Partitioning Clustering
- Expectation-Maximization

Attribute Importance



- Minimum Description Length
- Unsupervised pair-wise KL div.

Anomaly Detection



- 1 Class Support Vector Machine

Time Series

- Single & Double Exp. Smoothing

Predictive Queries

- Clustering
- Regression
- Anomaly Detection
- Feature Extraction

Feature Extraction & Creation

- Nonnegative Matrix Factorization
- Principal Component Analysis
- Singular Value Decomposition

Market Basket Analysis

- Apriori – Association Rules

Open Source R Algorithms

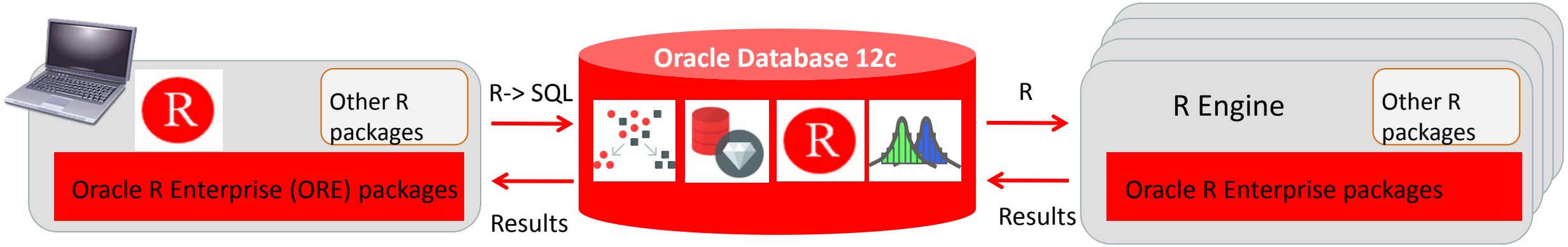
- Ability to run any R package via Embedded R mode

* supports partitioned models, text mining

ORACLE

Oracle Advanced Analytics

How Oracle R Enterprise Compute Engines Work



1 R-> SQL Transparency "Push-Down"

- R language for interaction with the database
- R-SQL Transparency Framework overloads R functions for scalable in-database execution
- Function overload for data selection, manipulation and transforms
- Interactive display of graphical results and flow control as in standard R
- Submit user-defined R functions for execution at database server under control of Oracle Database

2 In-Database Adv Analytical SQL Functions

- 15+ Powerful data mining algorithms (regression, clustering, AR, DT, etc._
- Run Oracle Data Mining SQL data mining functioning (ORE.odmSVM, ORE.odmDT, etc.)
- Speak "R" but executes as proprietary in-database SQL functions—machine learning algorithms and statistical functions
- Leverage database strengths: SQL parallelism, scale to large datasets, security
- Access big data in Database and Hadoop via SQL, R, and Big Data SQL

3 Embedded R Package Callouts

- R Engine(s) spawned by Oracle DB for database-managed parallelism
- ore.groupApply high performance scoring
- Efficient data transfer to spawned R engines
- Emulate map-reduce style algorithms and applications
- Enables production deployment and automated execution of R scripts

ORACLE

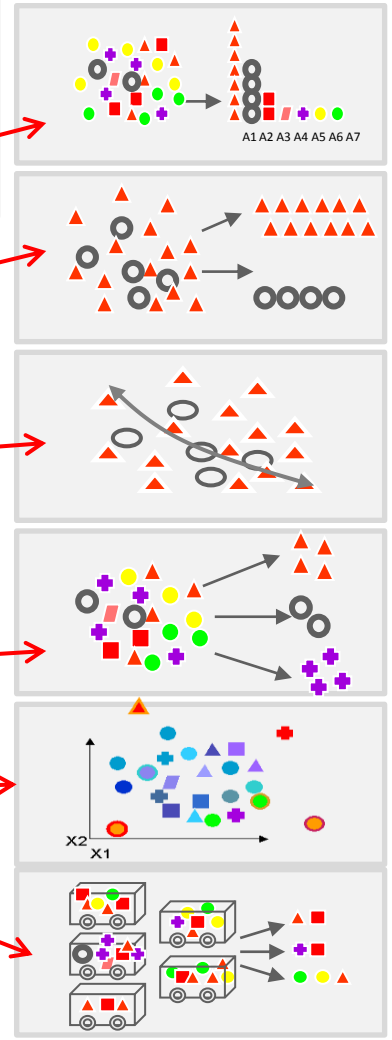
Data Mining & Anomaly Detection Concepts

ORACLE

What is Data Mining & Predictive Analytics?

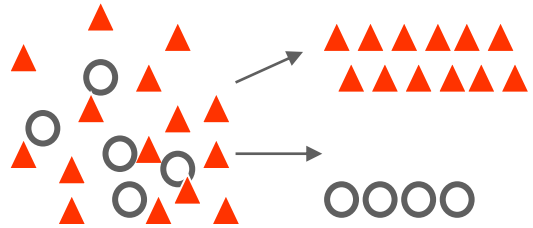
Automatically sifting through large amounts of data to create models that find previously hidden patterns, discover valuable new insights and make predictions

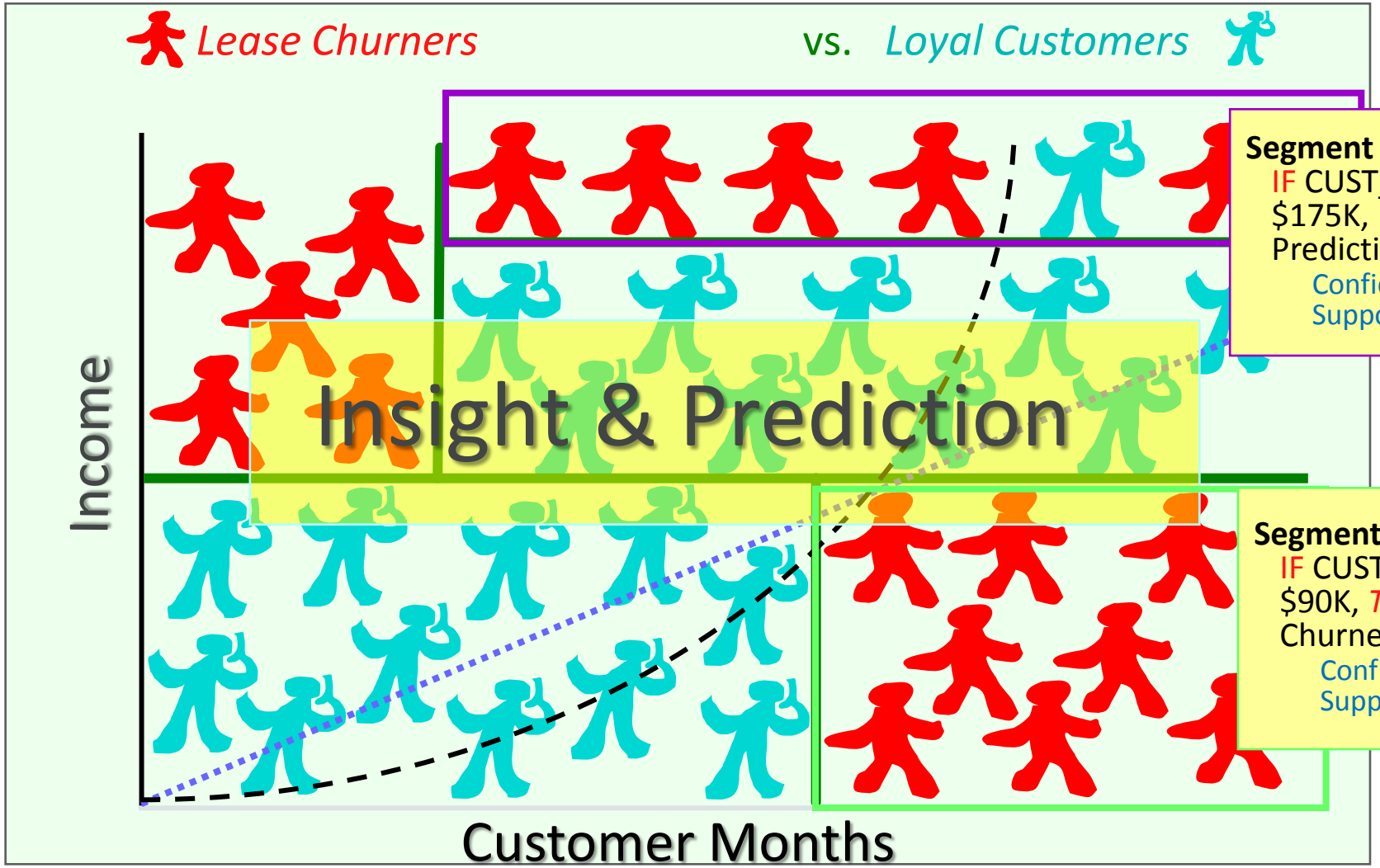
- Identify most important factor *(Attribute Importance)*
- Predict customer behavior *(Classification)*
- Predict or estimate a value *(Regression)*
- Find profiles of targeted people or items *(Decision Trees)*
- Segment a population *(Clustering)*
- Find fraudulent or “rare events” *(Anomaly Detection)*
- Determine co-occurring items in a “baskets” *(Associations)*



Data Mining Provides

Better Information, Valuable Insights and Predictions





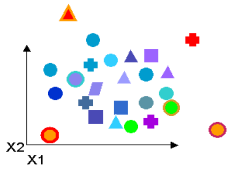
Segment #3
IF CUST_MO > 7 AND INCOME < $175K, *THEN*
Prediction = Lease Churner,
Confidence = 83%
Support = 6/39

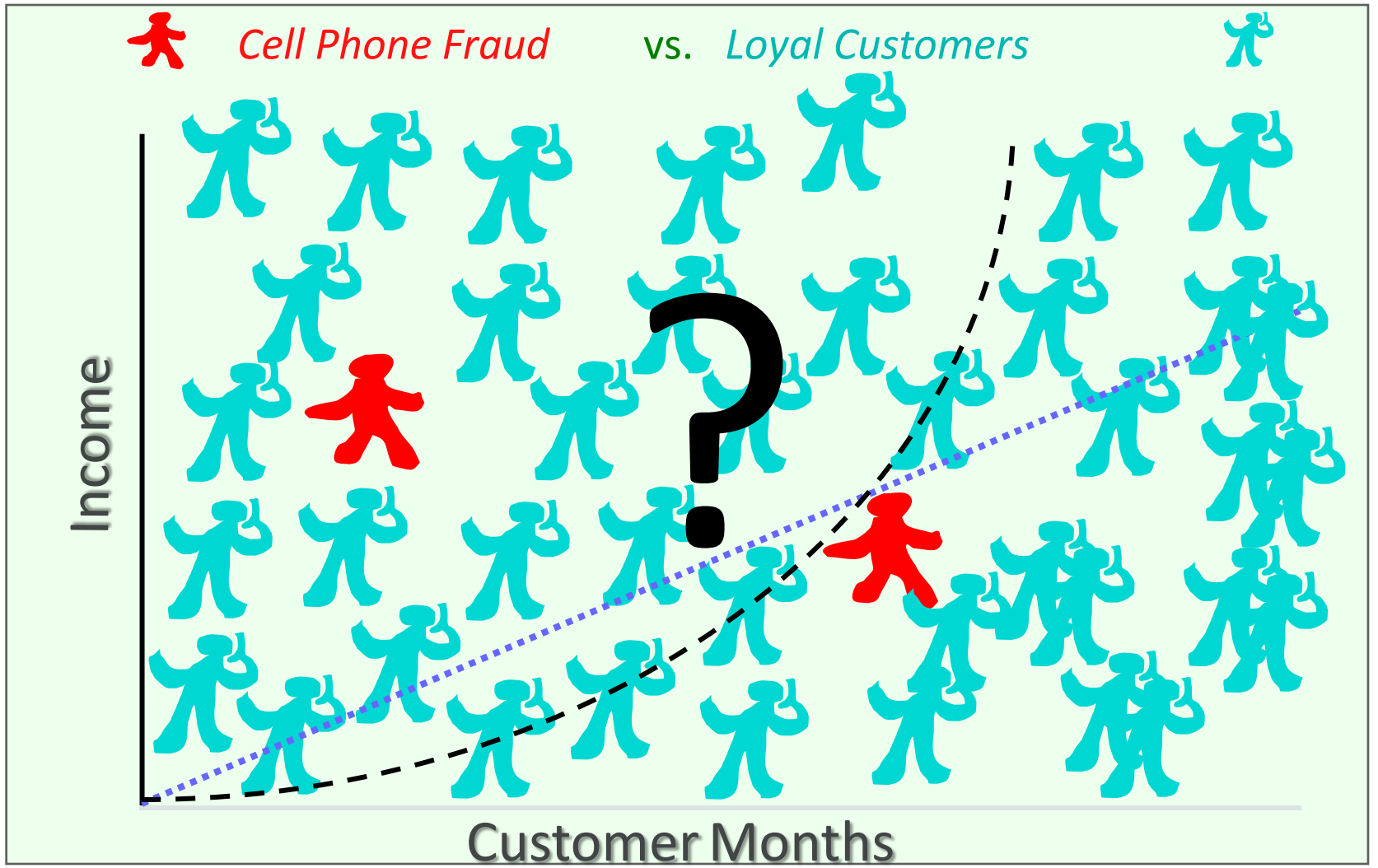
Segment #1
IF CUST_MO > 14 AND INCOME < $90K, *THEN* Prediction = Lease Churner
Confidence = 100%
Support = 8/39

Source: Inspired from *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management* by Michael J. A. Berry, Gordon S. Linoff

Data Mining When Lack Examples

Better Information, Valuable Insights and Predictions



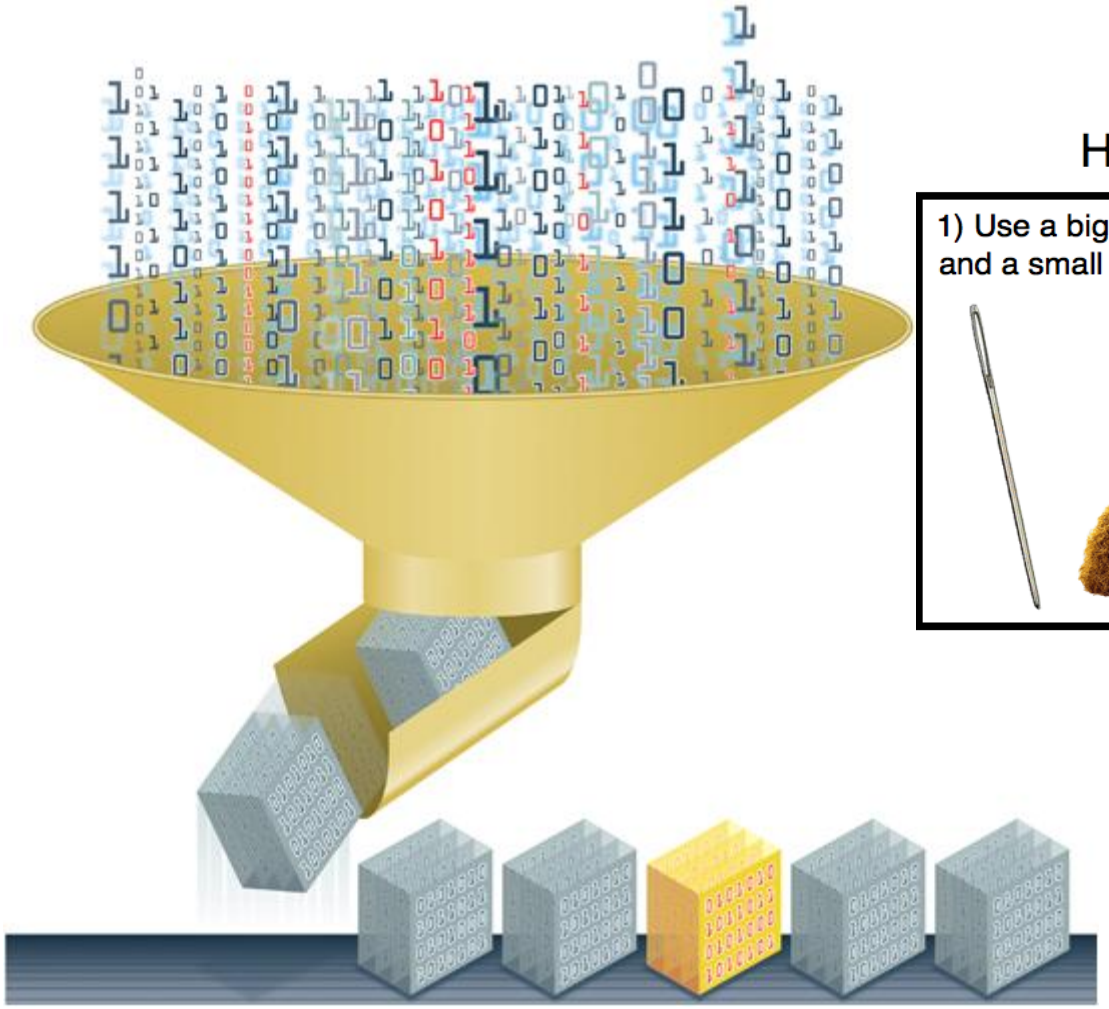


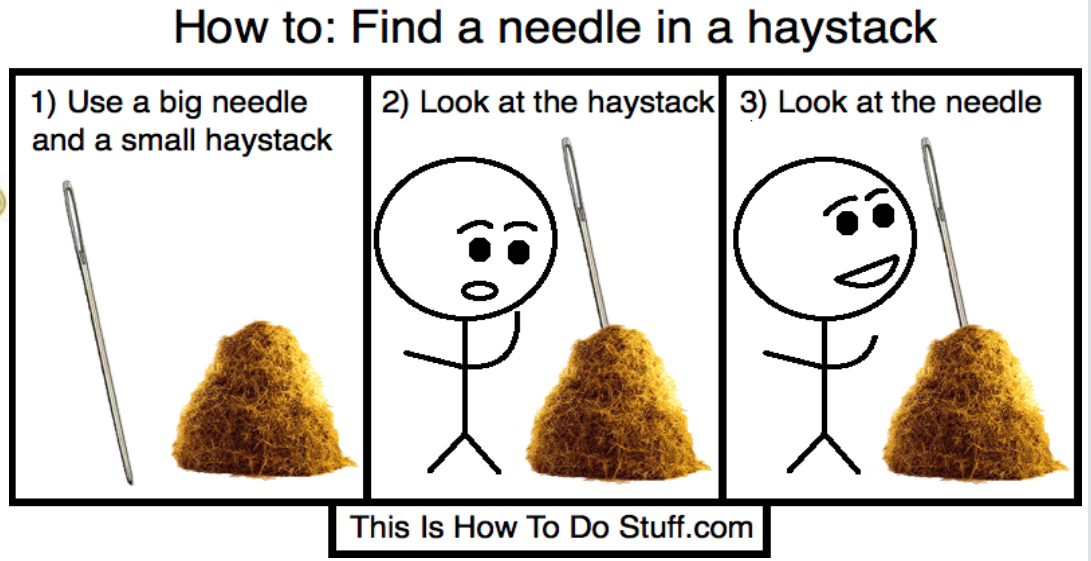
Source: Inspired from *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management* by Michael J. A. Berry, Gordon S. Linoff

Predictive Analytics & Data Mining

Finding Rare and Unusual Records in Large Datasets

- Finding needles in haystacks.
- Look for what's different….



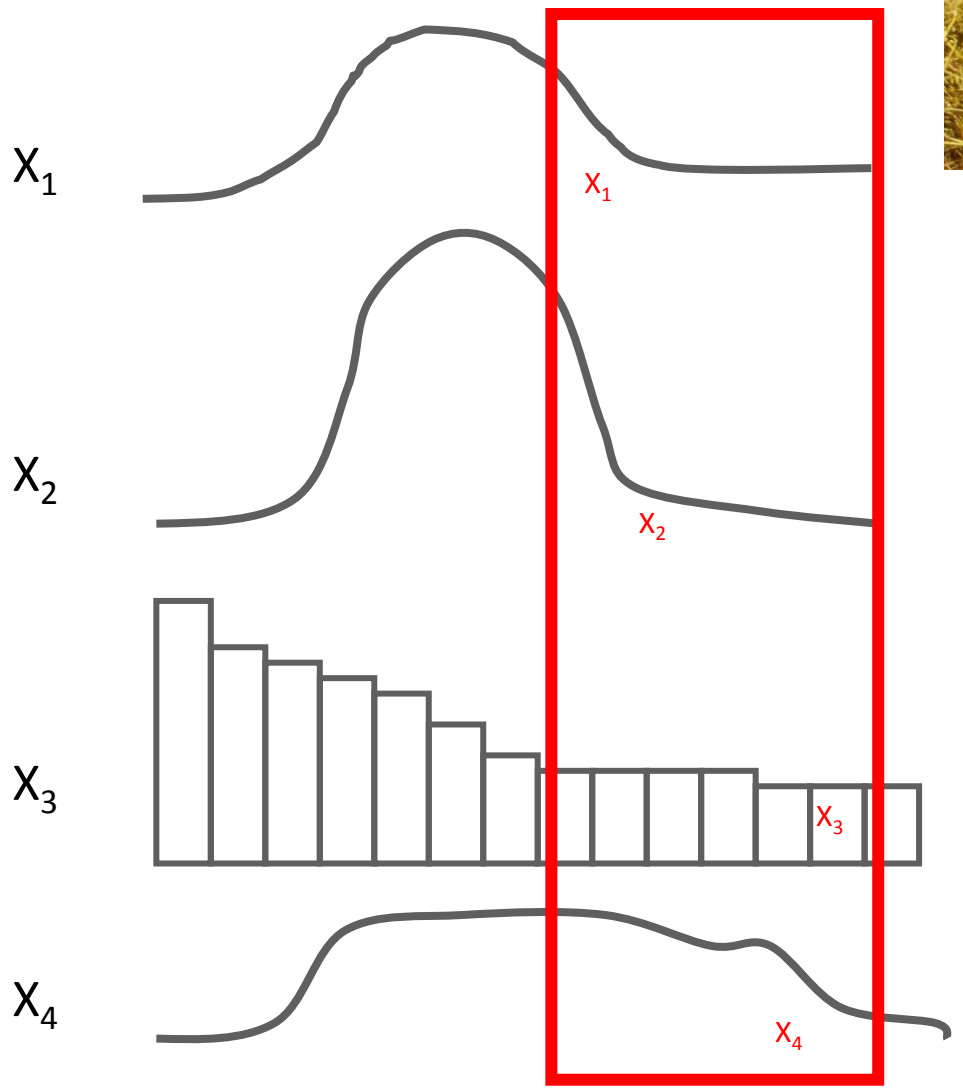


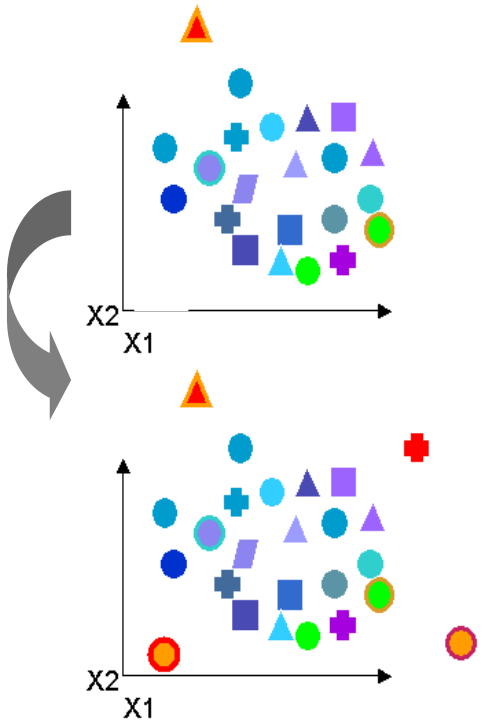
https://masonresearch.gmu.edu/2013/03/researcher-unlocks-the-big-potential-of-big-data/

ORACLE

Challenge: Finding Anomalies

- Considering multiple attributes
- Taken alone, may seem “normal”
- Taken collectively, a record may appear to be anomalous
- Look for what is “*different*”





A Real Fraud Example

My credit card statement—Can you see the fraud?

	May 22	1:14 PM	FOOD	Monaco Café	$127.38
	May 22	7:32 PM	WINE	Wine Bistro	$28.00
	…				
→	June 14	2:05 PM	MISC	Mobil Mart	$75.00
→	June 14	2:06 PM	MISC	Mobil Mart	$75.00
→	June 15	11:48 AM	MISC	Mobil Mart	$75.00
→	June 15	11:49 AM	MISC	Mobil Mart	$75.00
	May 28	6:31 PM	WINE	Acton Shop	$31.00
	May 29	8:39 PM	FOOD	Crossroads	$128.14
→	June 16	11:48 AM	MISC	Mobil Mart	$75.00
→	June 16	11:49 AM	MISC	Mobil Mart	$75.00

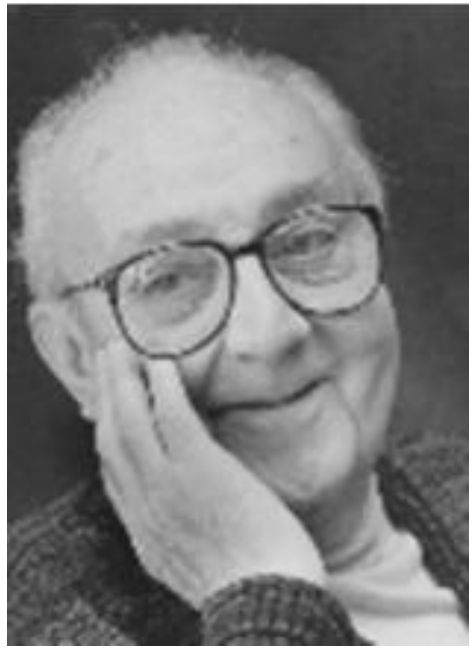
Total purchases exceeds time period average

Gas Station?

Monaco?

Pairs of $75?

All same $75 amount?



“Essentially, all models are wrong, …but some are useful.”

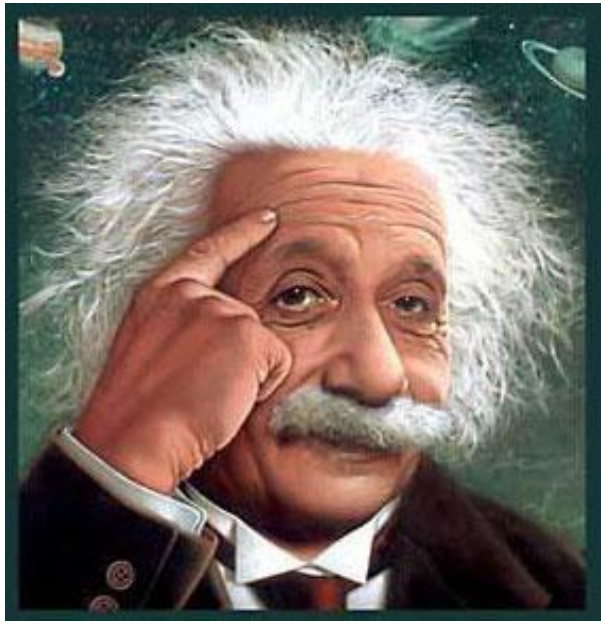
– George Box
(One of the most influential statisticians of the 20th century and a pioneer in the areas of quality control, time series analysis, design of experiments and Bayesian inference.)

Start with a Business Problem Statement

Clearly Define Problem

"If I had an hour to solve a problem I'd spend 55 minutes thinking about the problem and 5 minutes thinking about solutions."

— Albert Einstein

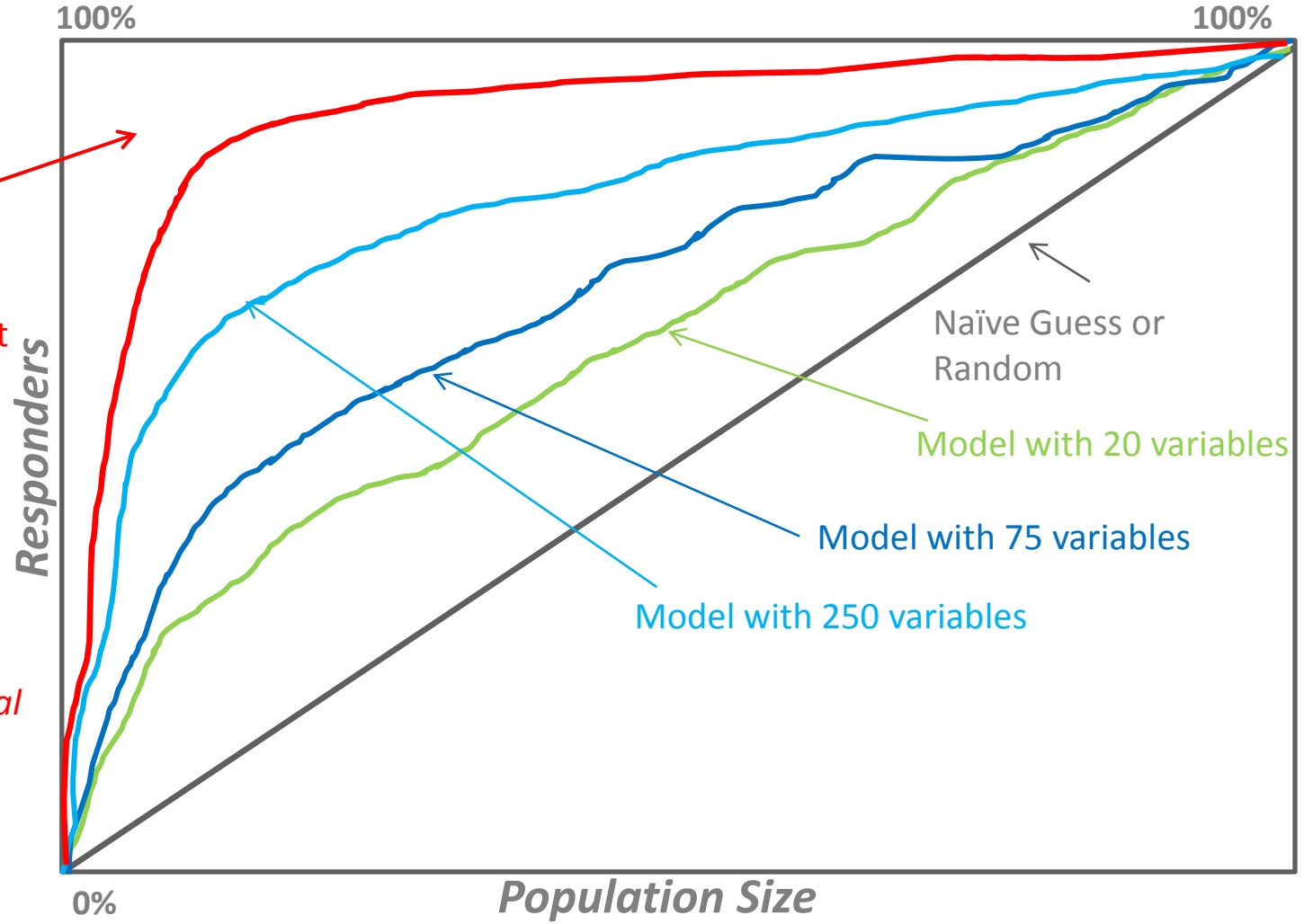


More Data Variety—Better Predictive Models

- Increasing sources of relevant data can boost model accuracy

Model with *"Big Data"* and hundreds -- thousands of input variables including:

- *Demographic data*
- *Purchase POS transactional data*
- *"Unstructured data", text & comments*
- *Spatial location data*
- *Long term vs. recent historical behavior*
- *Web visits*
- *Sensor data*
- *etc.*



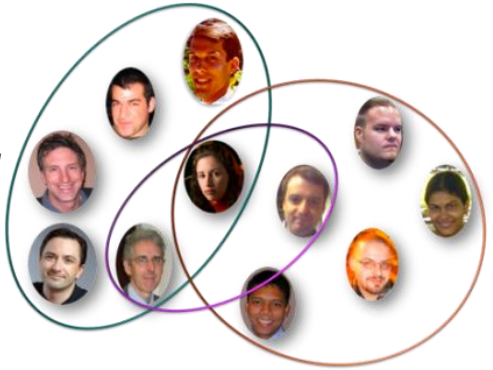
ORACLE

Multiple Data Sources/Types with Predictive Modeling

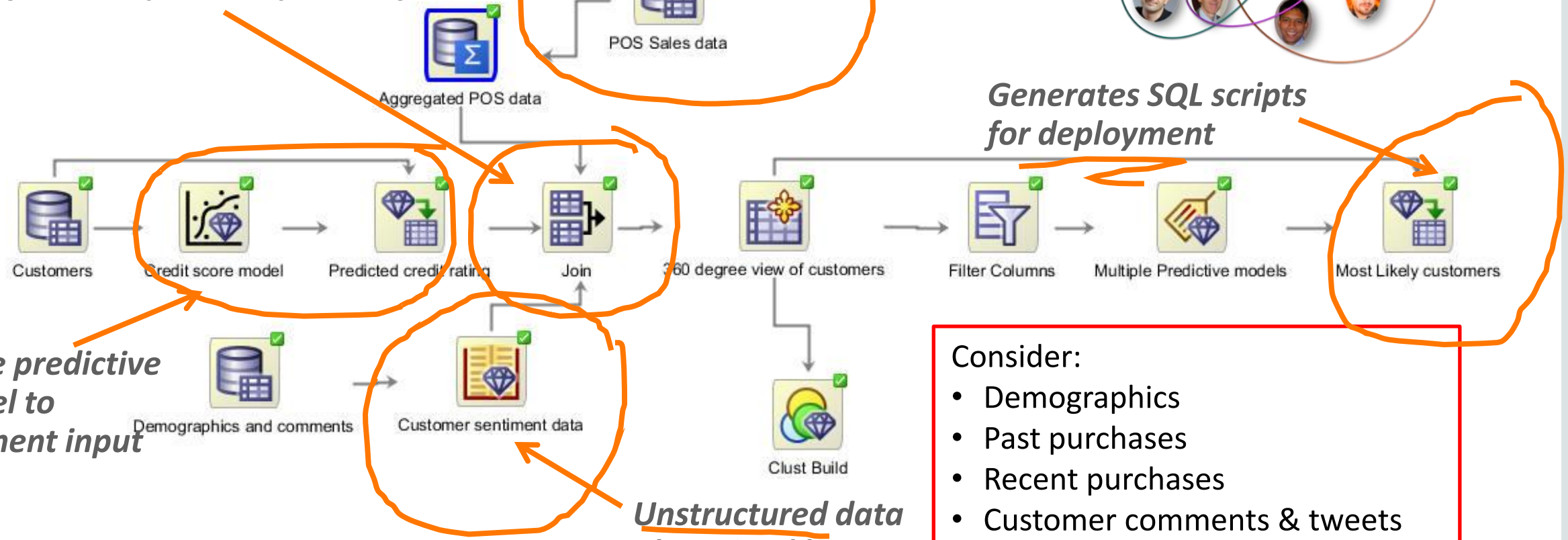
Ease of Deployment through SQL Script Generation

SQL Joins and arbitrary SQL transforms & queries – power of SQL

Transactional POS data



Generates SQL scripts for deployment



Inline predictive model to augment input data

Unstructured data also mined by algorithms

Consider:

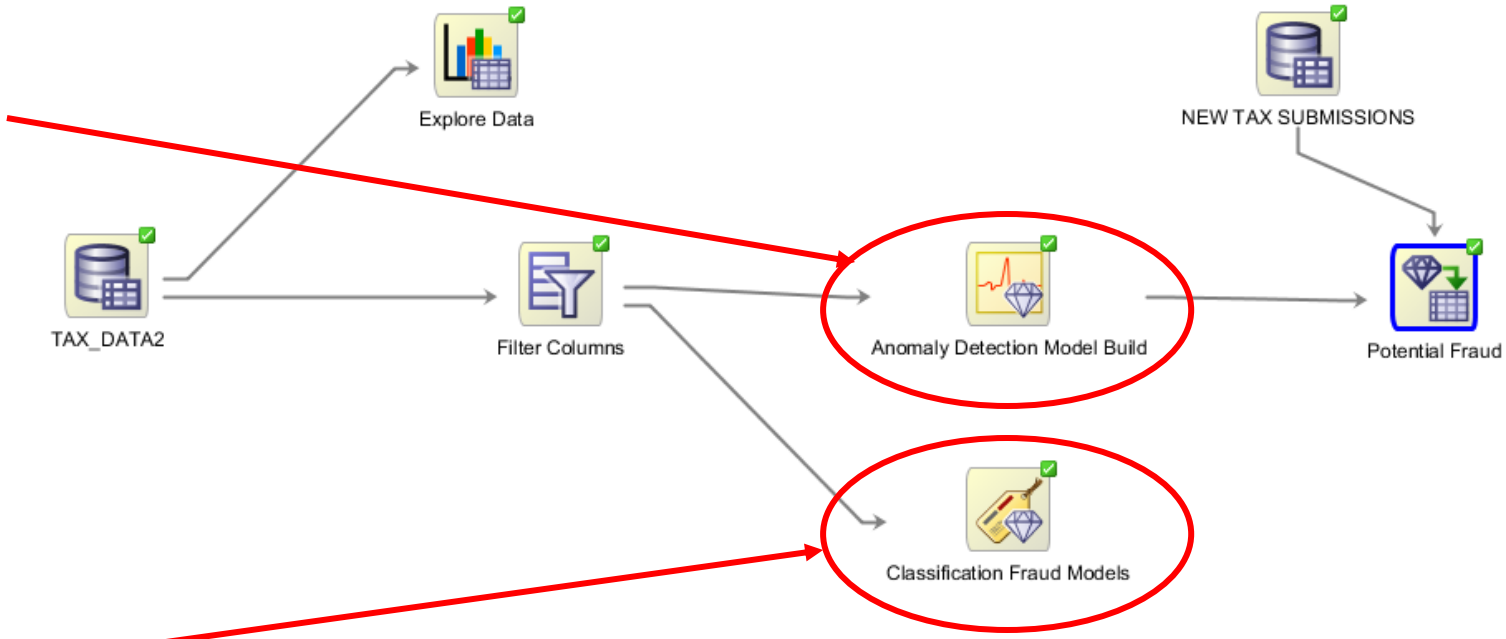
- Demographics
- Past purchases
- Recent purchases
- Customer comments & tweets

ORACLE

Tax Noncomplaince Audit Selection

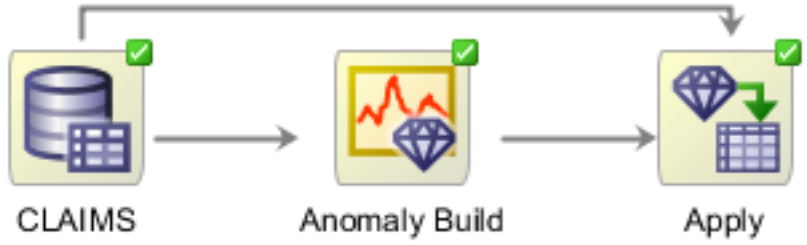
Two Example Approaches - There are many possible more!

- Anomaly Detection
  - Build 1-Class Support Vector Machine (SVM) models on “normal or compliant” tax submissions
    - Unsupervised machine learning when few know examples on which to train e.g. < 2%
  - Build Decision Tree models for classification of Noncompliant tax submissions (yes/no) based on historical 2011 data
    - Supervised machine learning approach when many known examples of target classes are available oh which to train

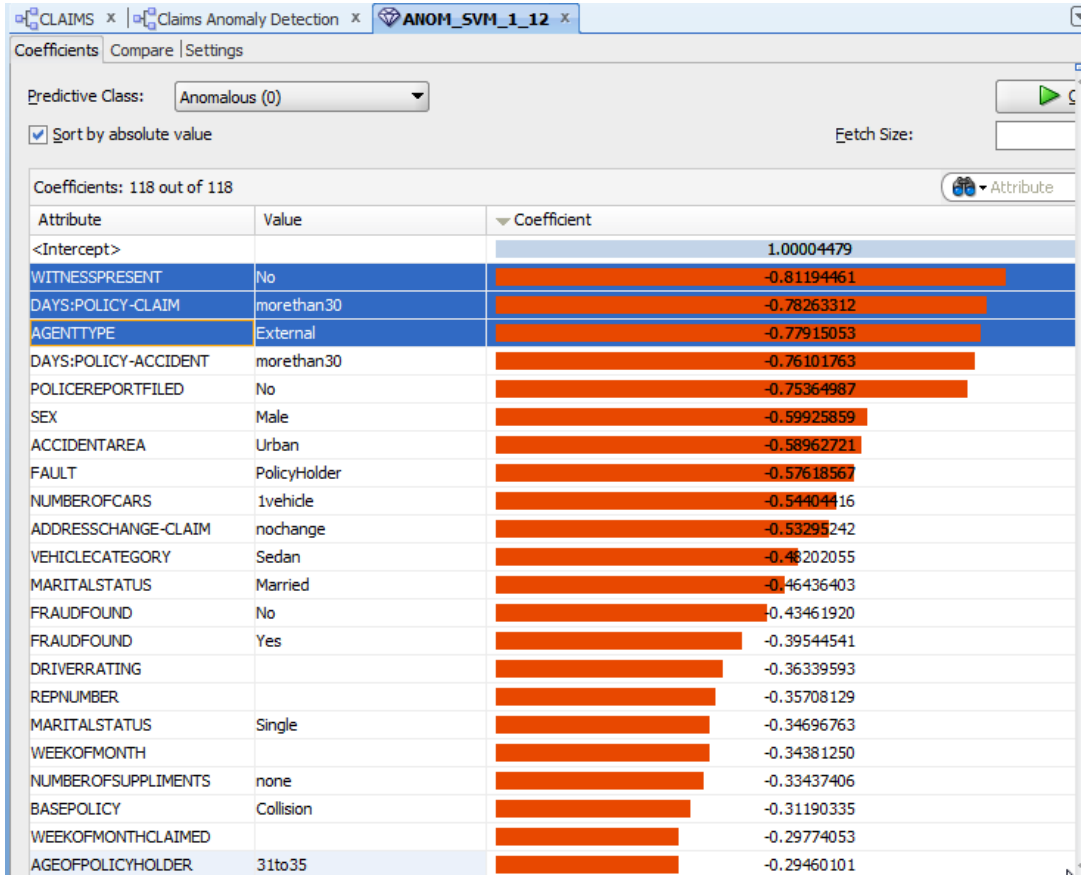


SQL Developer/Oracle Data Miner GUI

Anomaly Detection—Simple Conceptual Workflow



Train on “normal” records
Apply model and sort on likelihood to be “different”

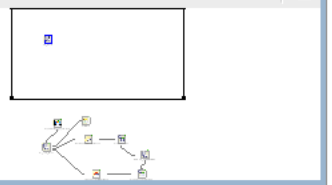


File Edit View Navigate Run Diagram Team Tools Window Help

Connections | Data Miner

- Connections
  - dmuser
    - ACME Mfg Paint Project
    - BERGERS R US
    - Chicago Crime
    - Customers R Us Project
    - Fighting Fraud
      - Expense Report Anomalies
      - Fun with fraud
      - *Tax Analytics*
        - 2nd Anomaly Detection Model
        - Clust Build
        - Noncompliant predictivemodels
        - 2011_TAX_DATA2
        - Explore Data grp_by Compaint
        - NEW TAX_SUBMISSIONS
        - Suspicious claims
        - Suspicious claims for review
        - TAX_DATA2
      - Text Mining etc
    - Healthcare R Us Project
    - IRIS dataset

Thumbnail



Tax Analytics - Structure | Reports

- Clust Build
- Explore Data grp_by Compaint
- Noncompliant predictivemodels
- 2011_TAX_DATA2
- NEW TAX_SUBMISSIONS
- 2nd Anomaly Detection Model
- Suspicious claims
- Suspicious claims for review
- TAX_DATA2
- Links

Start Page | dmuser9.sql | *Tax Analytics*

100% | Parallel Query Off

TAX_DATA2

TAX_DATA2 - Properties

Find

Data

Cache

Details

Source Table: DMUSER.TAX_DATA2

Data

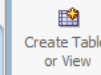
Name

Name	Alias	Data Type
ADDL_TAX_CREDIT		NUMBER
ADJUSTED_DEDUCTIONS		NUMBER
ADJUSTED_TAX_INC		NUMBER
AGE		NUMBER
ATI_BIN		VARCHAR2

Components

Workflow Editor

Data



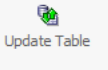
Create Table or View

Data Source

Explore Data

Graph

SQL Query



Update Table

Transforms

Text

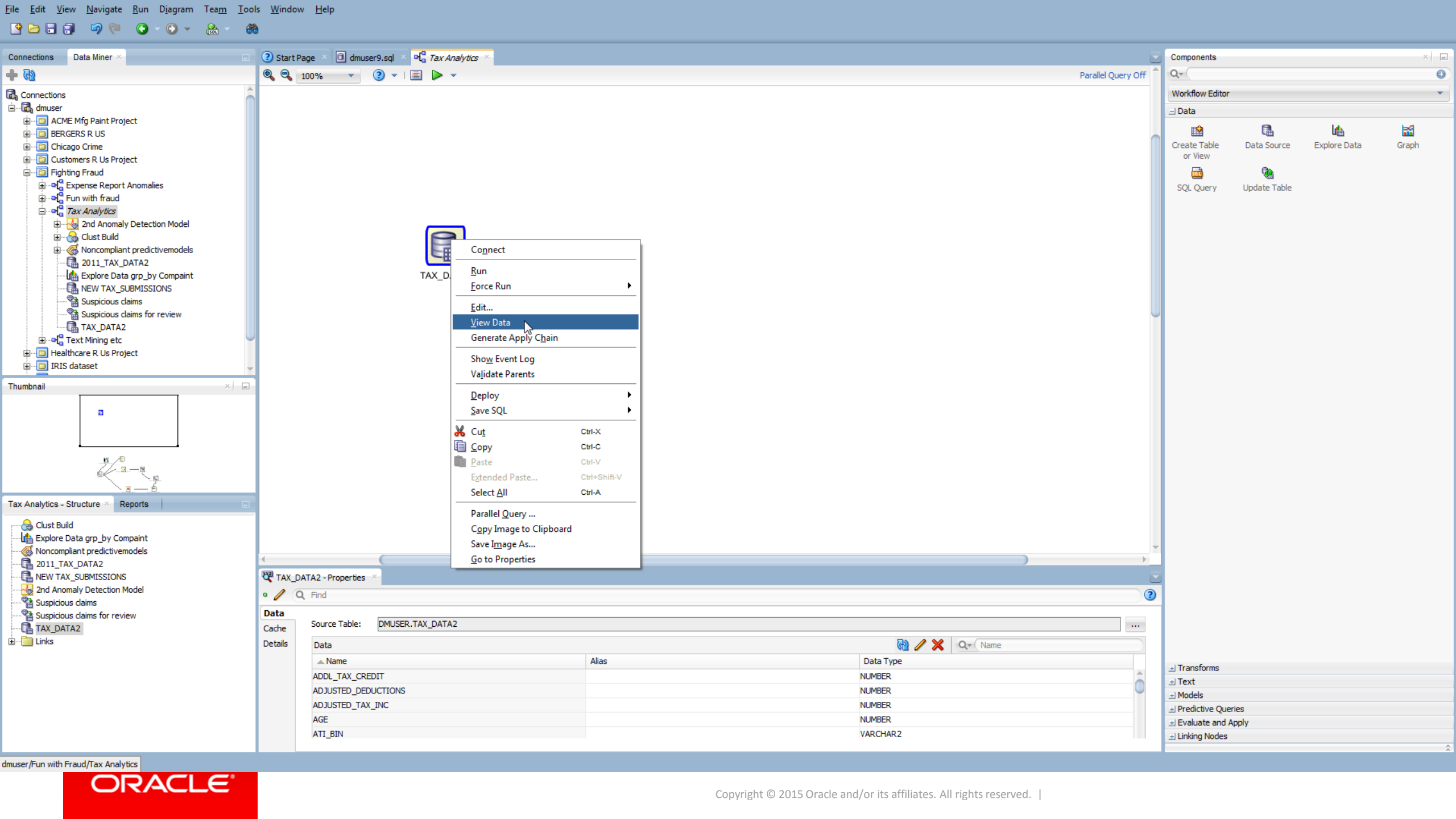
Models

Predictive Queries

Evaluate and Apply

Linking Nodes

dmuser/Fun with Fraud/Tax Analytics



File Edit View Navigate Run Team Tools Window Help

Connections | Data Miner

- Connections
  - dmuser
    - ACME Mfg Paint Project
    - BERGERS R US
    - Chicago Crime
    - Customers R Us Project
    - Fighting Fraud
      - Expense Report Anomalies
      - Fun with fraud
      - Tax Analytics
        - 2nd Anomaly Detection Model
        - Clust Build
        - Noncompliant predictivemodels
        - 2011_TAX_DATA2
        - Explore Data grp_by Compaint
        - NEW TAX_SUBMISSIONS
        - Suspicious claims
        - Suspicious claims for review
        - TAX_DATA2
      - Text Mining etc
    - Healthcare R Us Project
      - Brain tissue text mining
      - Hospitals R US
      - Lymphoma Patients Analytics
      - NCI60 Cancer Genes Prediction
      - NCI60 gene expression analysis
    - IRIS dataset
    - ONTIME Airline Data
    - Titanic Survivors

Tax Analytics - Structure | Reports

No Structure

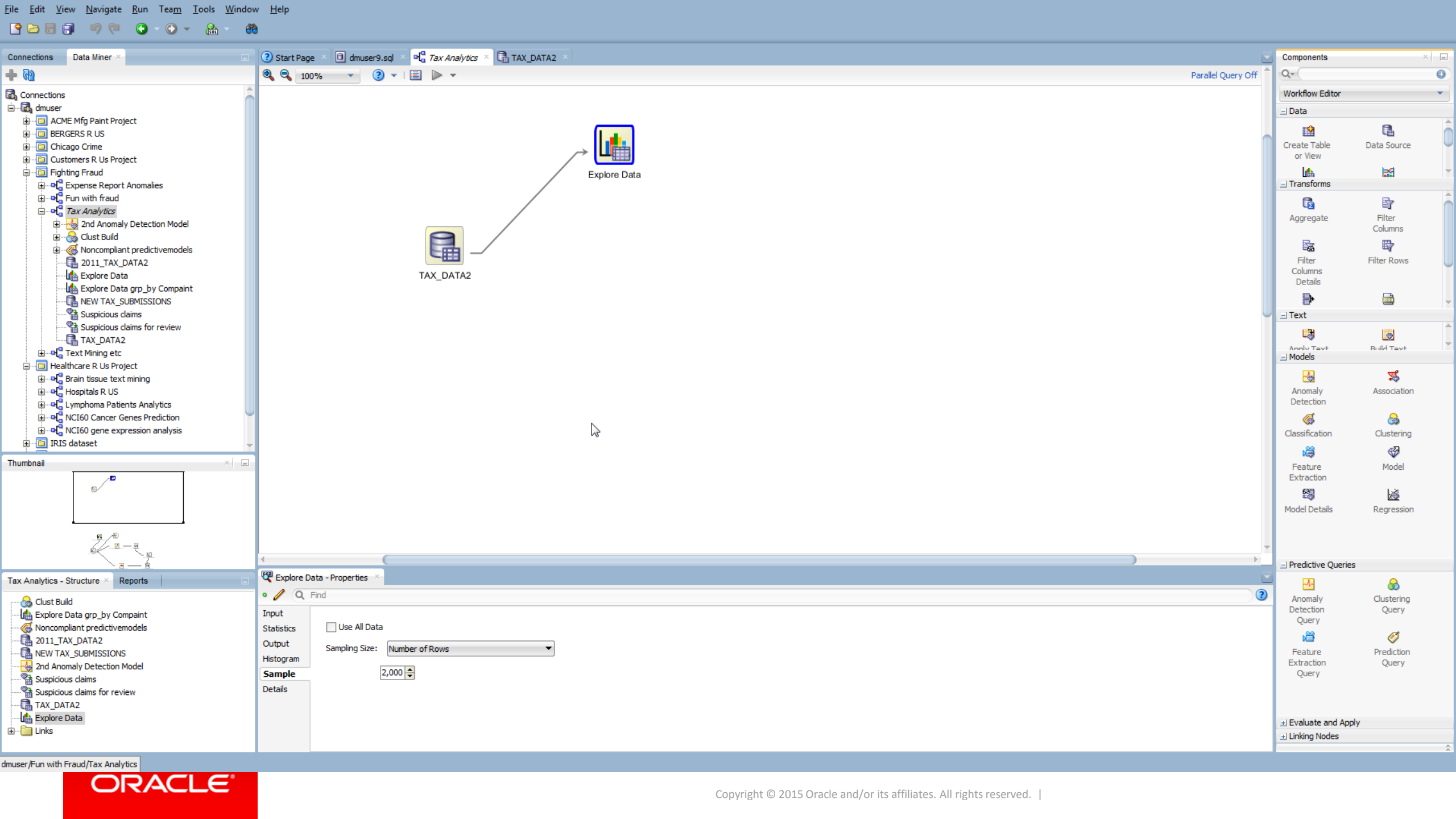
Start Page | dmuser9.sql | Tax Analytics | TAX_DATA2

Data | Columns | SQL

View: Actual Data | Sort... | Parallel Query Off... | Filter: Enter Where Clause

	TAX_PERSON_ID	COMPLIANT	SALARY	MARITAL_STATUS	MEDICAL_DEDUCTIONS	STATE	NUM_LATE_PAYMENTS	SEX	ADJUSTED_TAX_INC	ITEMIZED_DEDUCTIONS	ADDL_TAX_CREDIT	ADJUSTED_DEDUCTIONS	YRS_RESIDENCE	HOUSE_OWNERSHIP	MONTHLY_CHECKS_WRITTEN	N_MO
1	CU4642	No	61,150	DIVORCED	2,500	NY	0	F	17,888	0	0	25	5	1	0	
2	CU4643	Yes	64,019	DIVORCED	2,705	CA	5	M	31,005	2,550	0	254	2	1	3	
3	CU10089	No	62,850	MARRIED	1,000	OH	3	F	23,413	201	3,608	25	5	1	1	
4	CU10090	No	72,612	DIVORCED	3,650	NY	2	F	27,853	0	0	25	5	2	2	
5	CU10091	No	55,395	MARRIED	711	CA	3	M	21,849	450	0	25	3	1	5	
6	CU10092	No	58,475	MARRIED	1,500	NY	1	M	26,219	0	0	25	1	1	1	
7	CU10093	No	75,086	MARRIED	1,000	WA	5	F	24,772	0	0	25	1	1	1	
8	CU10094	No	62,924	DIVORCED	3,000	MO	2	F	24,131	0	0	1,185	1	1	6	
9	CU10095	Yes	73,369	MARRIED	600	CA	4	M	22,042	3,550	0	290	3	1	3	
10	CU10096	No	67,407	MARRIED	1,800	MI	3	M	21,852	0	0	2,828	3	1	0	
11	CU10097	Yes	61,502	WIDOWED	3,000	NY	6	F	23,776	6,300	0	25	1	1	2	
12	CU10099	No	67,533	MARRIED	5,300	NY	0	F	24,383	0	0	25	1	1	0	
13	CU10100	Yes	64,427	MARRIED	1,100	OR	4	F	33,307	950	0	25	4	2	7	
14	CU7549	No	61,860	MARRIED	1,005	NY	3	M	23,965	201	0	241	3	1	5	
15	CU7550	No	59,081	SINGLE	530	CA	3	M	17,470	700	0	540	3	1	6	
16	CU9766	No	69,518	DIVORCED	4,850	MI	0	M	30,280	0	0	25	1	1	0	
17	CU9342	No	69,480	MARRIED	11,717	CA	1	M	27,970	0	0	25	2	1	1	
18	CU9343	No	62,059	SINGLE	0	MI	2	M	19,715	0	0	25	2	0	1	
19	CU9344	No	69,913	WIDOWED	5,000	CA	2	F	23,778	0	0	25	4	1	1	
20	CU9345	Yes	59,090	WIDOWED	5,000	CA	5	F	34,473	25,000	32,391	4,166	1	2	17	
21	CU9347	No	69,457	MARRIED	773	CA	4	F	33,064	501	0	25	1	1	1	
22	CU9348	Yes	67,403	SINGLE	0	MI	2	M	9,551	750	0	1,014	3	0	2	
23	CU9350	No	59,081	MARRIED	1,515	FL	0	M	24,270	0	0	25	4	1	0	
24	CU9351	No	68,746	MARRIED	1,500	MN	2	F	25,087	7,400	0	2,168	1	1	1	
25	CU9352	Yes	55,530	DIVORCED	1,500	CA	5	F	20,183	4,400	0	2,124	1	1	2	
26	CU9353	No	73,182	MARRIED	6,000	NY	6	F	35,596	0	0	25	1	2	0	
27	CU9354	No	74,534	MARRIED	1,000	CA	3	M	21,834	0	0	25	3	1	9	
28	CU9355	Yes	63,099	DIVORCED	500	NY	4	M	19,575	2,500	0	25	3	1	12	
29	CU9356	Yes	67,719	DIVORCED	750	CA	6	M	23,230	4,300	0	25	3	1	3	
30	CU9357	No	74,473	SINGLE	0	NY	0	M	21,918	0	0	25	2	0	1	
31	CU9358	Yes	68,064	DIVORCED	550	MN	6	F	24,816	4,350	0	25	1	1	1	
32	CU9360	No	69,247	SINGLE	0	CA	0	M	24,812	0	0	25	4	0	1	
33	CU9362	No	64,355	SINGLE	0	MI	6	M	12,989	0	56,607	25	4	0	0	
34	CU9364	Yes	66,157	DIVORCED	850	FL	5	M	24,939	5,050	0	1,535	1	1	3	
35	CU6874	No	64,705	SINGLE	0	NY	0	M	15,976	0	0	25	2	0	0	
36	CU6875	No	67,859	MARRIED	800	UT	4	M	28,965	700	0	25	4	2	2	
37	CU6876	No	66,503	MARRIED	900	NY	3	M	34,726	7,200	10,976	1,629	4	2	5	
38	CU6878	No	97,455	WIDOWED	10,000	MI	2	F	27,464	0	0	25	5	1	0	
39	CU6880	No	59,340	DIVORCED	3,200	MI	2	M	24,435	14,250	0	5,330	2	1	2	
40	CU6881	No	68,641	DIVORCED	2,000	NY	2	M	27,760	4,400	0	147	2	1	1	

dmuser/Fun with Fraud/Tax Analytics



File Edit View Navigate Run Team Tools Window Help

Connections | Data Miner

- Connections
  - dmuser
    - ACME Mfg Paint Project
    - BERGERS R US
    - Chicago Crime
    - Customers R Us Project
    - Fighting Fraud
      - Expense Report Anomalies
      - Fun with fraud
      - Tax Analytics
        - 2nd Anomaly Detection Model
        - Clust Build
        - Noncompliant predictivemodels
        - 2011_TAX_DATA2
        - Explore Data
        - Explore Data grp_by Compaint
        - NEW TAX_SUBMISSIONS
        - Suspicious claims
        - Suspicious claims for review
        - TAX_DATA2
      - Text Mining etc
    - Healthcare R Us Project
      - Brain tissue text mining
      - Hospitals R US
      - Lymphoma Patients Analytics
      - NCI60 Cancer Genes Prediction
      - NCI60 gene expression analysis
    - IRIS dataset
    - ONTIME Airline Data
    - Titanic Survivors

Explore Data - Structure | Reports

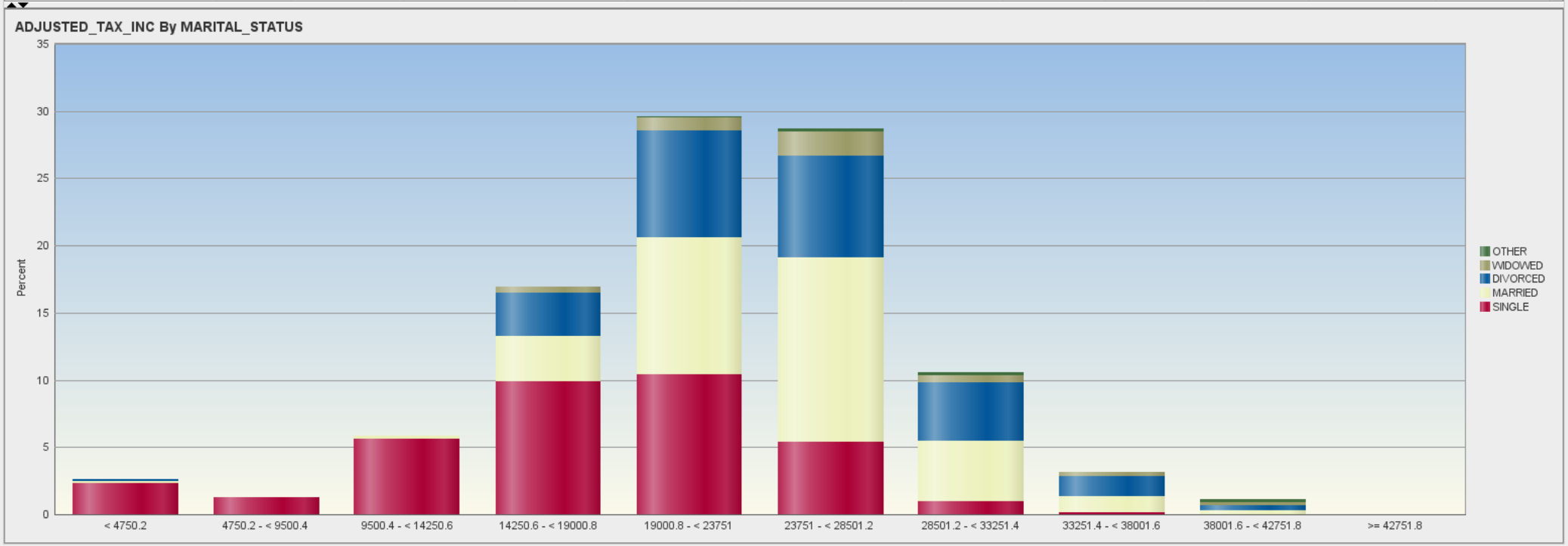
No Structure

Start Page | dmuser9.sql | Tax Analytics | TAX_DATA2 | Explore Data

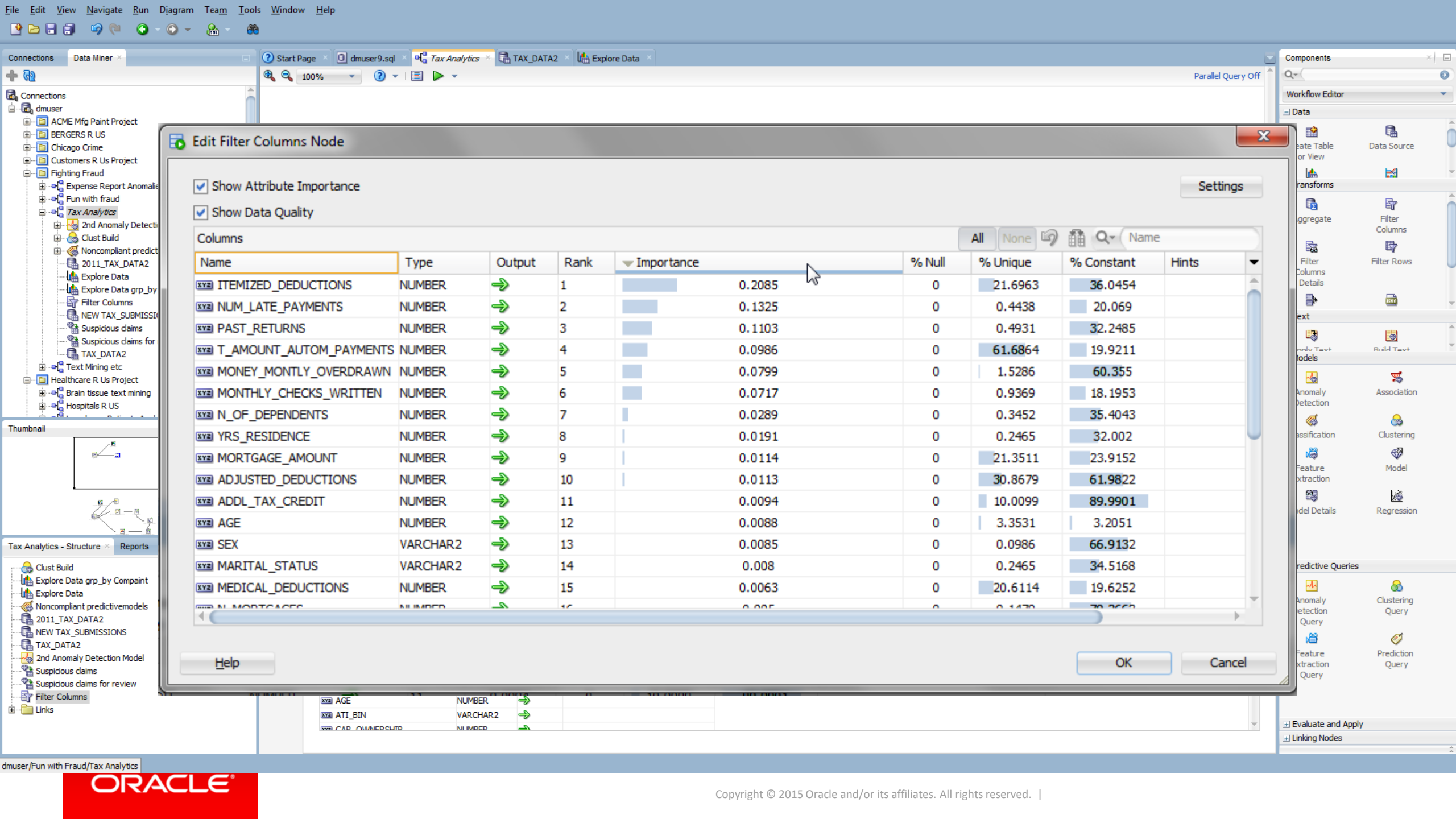
Statistics | Data | Columns | SQL

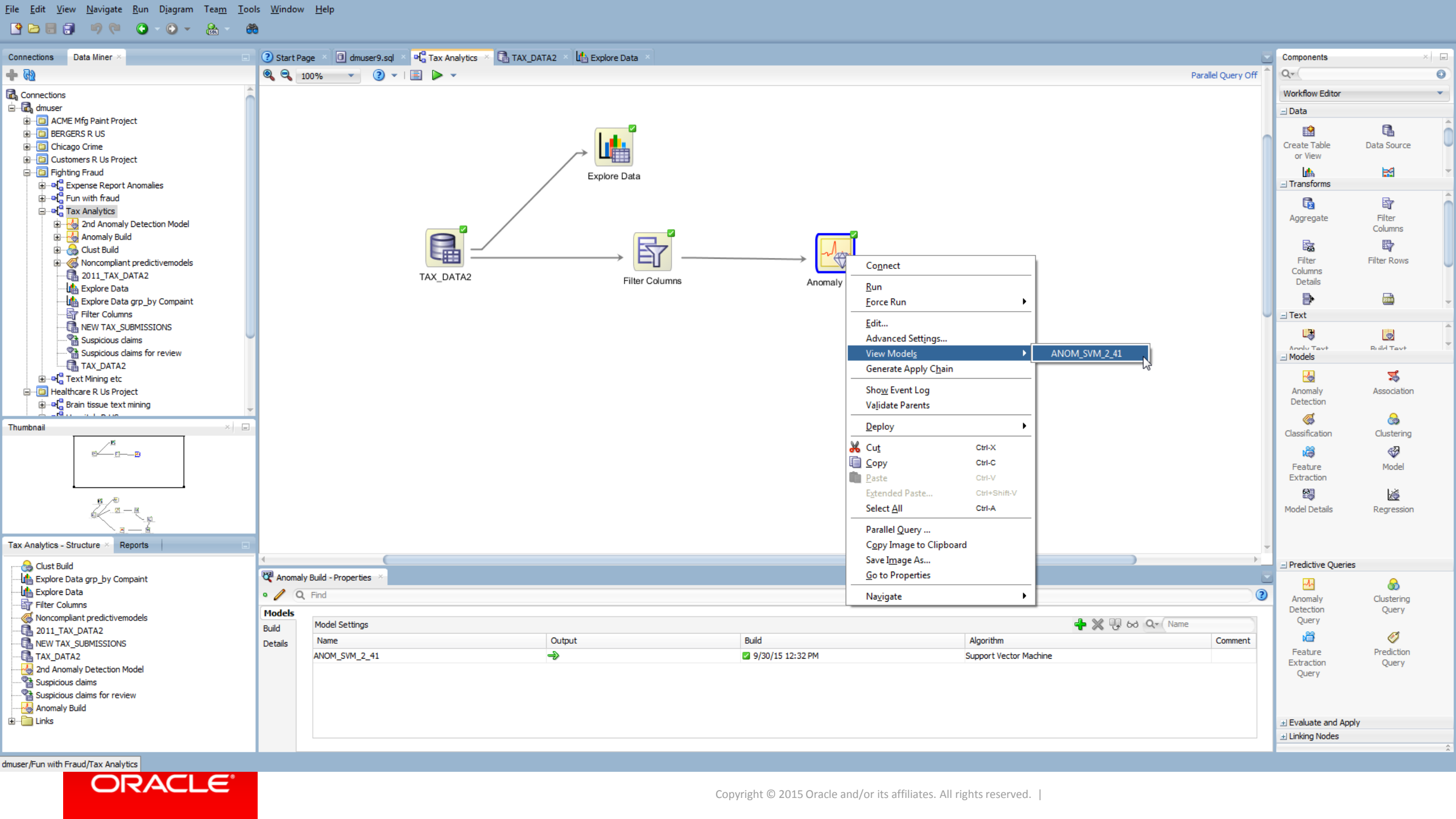
Statistics: 10 Columns from 2,005 Rows(Sampled) Group by: MARITAL_STATUS Show Nulls Filter: Name

Name	Histogram	Data Type	Percent NULLs	Distinct Values	Distinct Percent	Mode	Average	Median	Min Value	Max Value	Standard Deviation	Variance
ADDL_TAX_CREDIT		NUMBER	0	201	10.0249		3,248.9596	0	0	608,951	21,652.8888	468,847,592.913
ADJUSTED_DEDUCTIONS		NUMBER	0	616	30.7232		1,110.3786	25	25	24,720	3,287.6536	10,808,666.0428
ADJUSTED_TAX_INC		NUMBER	0	1,852	92.3691		22,276.7042	23,054	0	47,502	6,889.3259	47,462,811.6485
AGE		NUMBER	0	65	3.2419		36.9985	36	0	84	14.2732	203.7231
ATI_BIN		VARCHAR2	0	4	0.1995	HIGH						
CAR_OWNERSHIP		NUMBER	0	2	0.0998		0.9392	1	0	1	0.2391	0.0572
COMPLIANT		VARCHAR2	0	2	0.0998	No						
CREDIT_CARD_LIMITS		NUMBER	0	23	1.1471		1,234.2643	1,000	500	5,000	806.8432	650,996.0279
FIRST		VARCHAR2	0	1,391	69.3766	<Other>						
HAS_CHILDREN		NUMBER	0	2	0.0998		0.4983	0	0	1	0.5001	0.2501
HOUSE_OWNERSHIP		NUMBER	0	3	0.1496		0.8095	1	0	2	0.5178	0.2681



dmuser/Fun with Fraud/Tax Analytics





File Edit View Navigate Run Team Tools Window Help

Connections | Data Miner

- Connections
  - dmuser
    - ACME Mfg Paint Project
    - BERGERS R US
    - Chicago Crime
    - Customers R Us Project
    - Fighting Fraud
      - Expense Report Anomalies
      - Fun with fraud
      - Tax Analytics
        - 2nd Anomaly Detection Model
        - Anomaly Build
        - Clust Build
        - Noncompliant predictivemodels
        - 2011_TAX_DATA2
        - Explore Data
        - Explore Data grp_by Compaint
        - Filter Columns
        - NEW TAX_SUBMISSIONS
        - Suspicious claims
        - Suspicious claims for review
        - TAX_DATA2
      - Text Mining etc
    - Healthcare R Us Project
      - Brain tissue text mining
      - Hospitals R US
      - Lymphoma Patients Analytics
      - NCI60 Cancer Genes Prediction
      - NCI60 gene expression analysis
    - IRIS dataset

ANOM_SVM_2_41 - Structure | Reports

No Structure

Start Page | dmuser9.sql | Tax Analytics | TAX_DATA2 | Explore Data | ANOM_SVM_2_41

Coefficients | Compare | Settings

Predictive Class: Anomalous (0) ☑ Sort by absolute value

Query

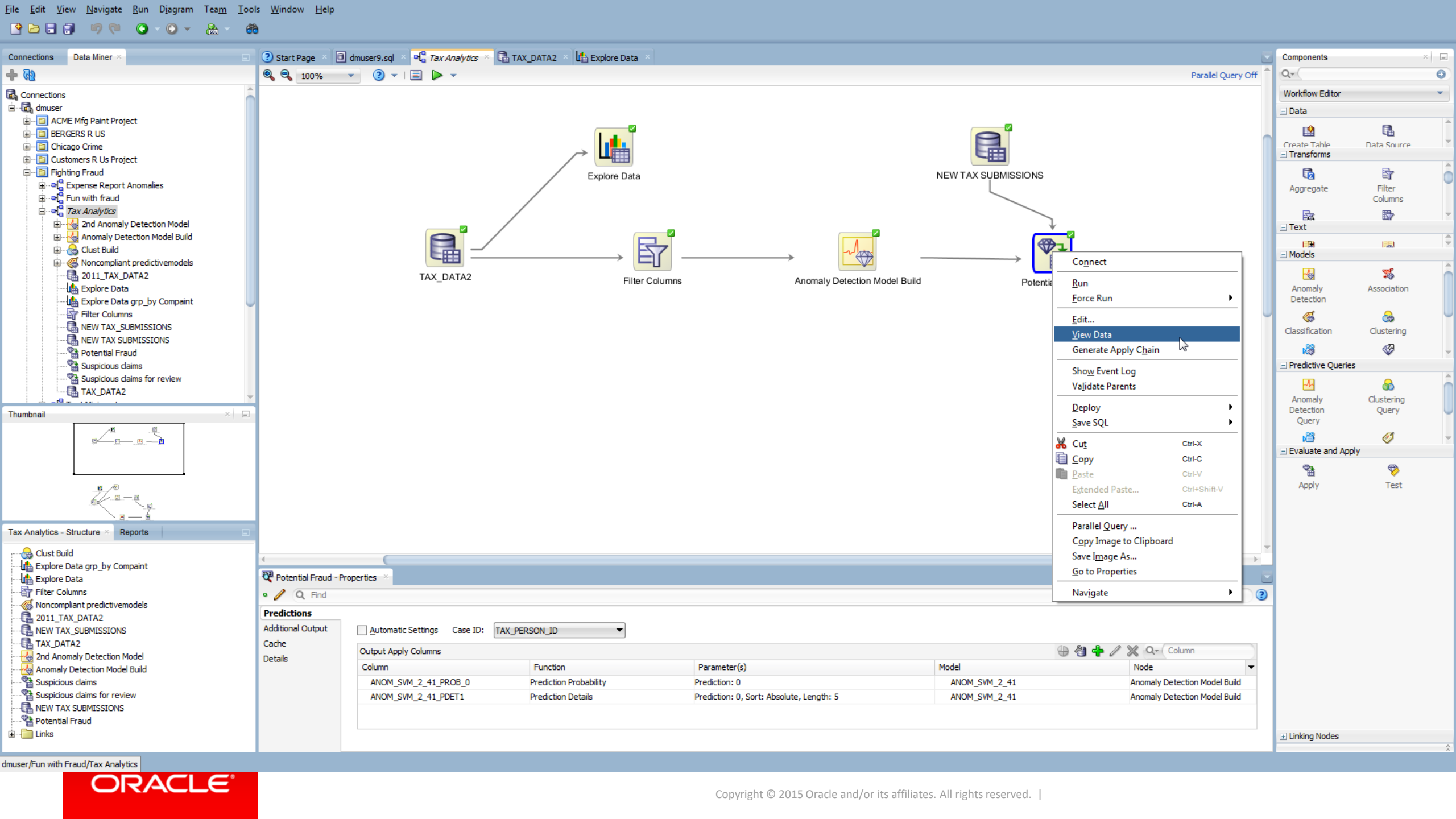
Fetch Size: 1,000

Coefficients: 163 out of 163

Attribute

Attribute	Value	Coefficient
MONEY_MONTLY_OVERDRAWN		-1.13904008
<Intercept>		0.99943098
CAR_OWNERSHIP	1	-0.93255699
COMPLIANT	No	-0.81425540
SEX	M	-0.74711325
HAS_CHILDREN	0	-0.70319249
HAS_CHILDREN	1	-0.69995852
SEX	F	-0.65603776
ADJUSTED_TAX_INC		-0.60285659
COMPLIANT	Yes	-0.58889561
MARITAL_STATUS	SINGLE	-0.53756055
HOUSE_OWNERSHIP	0	-0.53289224
N_MORTGAGES	0	-0.53289224
AGE		-0.53040870
HOUSE_OWNERSHIP	1	-0.49951563
N_MORTGAGES	1	-0.49951563
SALARY		-0.49949007
CAR_OWNERSHIP	0	-0.47059402
N_OF_DEPENDENTS		-0.45403275
NUM_LATE_PAYMENTS		-0.42053814
ATI_BIN	MEDIUM	-0.39325946
HOUSE_OWNERSHIP	2	-0.37074313
N_MORTGAGES	2	-0.37074313
YRS_RESIDENCE	1	-0.36680397
ATI_BIN	LOW	-0.35603616
YRS_RESIDENCE	2	-0.35204979
ATI_BIN	HIGH	-0.34842313
MARITAL_STATUS	MARRIED	-0.32720258
REGION	West	-0.32304288
MARITAL_STATUS	DIVORCED	-0.31805017
REGION	Midwest	-0.30989397
ATI_BIN	VERY HIGH	-0.30543226
REGION	NorthEast	-0.28570493
YRS_RESIDENCE	3	-0.27682231
CREDIT_CARD_LIMITS		-0.26067944
NUM_HEALTH_DEP		-0.25663882
REGION	South	-0.25520768
MONTHLY_CHECKS_WRITTEN		-0.25053403
PAST_RETURNS		-0.24792877
YRS_RESIDENCE	4	-0.22984737
REGION	Southwest	-0.22930155

dmuser/Fun with Fraud/Tax Analytics



File Edit View Navigate Run Team Tools Window Help

Connections | Data Miner

Connections
- dmuser
  - ACME Mfg Paint Project
  - BERGERS R US
  - Chicago Crime
  - Customers R Us Project
  - Fighting Fraud
    - Expense Report Anomalies
    - Fun with fraud
    - *Tax Analytics*
      - 2nd Anomaly Detectio
      - Anomaly Detection Mc
      - Clust Build
      - Noncompliant predicti
      - 2011_TAX_DATA2
      - Explore Data
      - Explore Data grp_by (
      - Filter Columns
      - NEW TAX_SUBMISSIO
      - NEW TAX SUBMISSIO
      - Potential Fraud
      - Suspicious claims
      - Suspicious claims for r
      - TAX_DATA2
    - Text Mining etc
  - Healthcare R Us Project
    - Brain tissue text mining
    - Hospitals R US
    - Lymphoma Patients Analy

Potential Fraud - Structure | Reports

No Structure

Start Page | dmuser9.sql | *Tax Analytics* | TAX_DATA2 | Explore Data | Potential Fraud

Data | Columns | SQL

View: Actual Data | Sort... | Parallel Query Off... | Filter: Enter Where Clause

	TAX_PERSON_ID	ANOM_SVM_2_41_PROB_0	PAST_RETURNS	ADJUSTED_TAX_INC	MEDICAL_DEDUCTIONS	ANOM_SVM_2_41_PDET1	ADJUSTED_DEDUCTIONS	CAR_OWNERSHIP	MARITAL_STATUS	LAST
1	CU5993	0.5893835406129992	1	26,551	8,000	<Details algorithm="Support Vector M...	143	1	WIDOWED	KIE...
2	CU5626	0.5603560735513573	2	0	2,275	<Details algorithm="Support Vector M...	62	1	DIVORCED	TR...
3	CU9114	0.5580233950203011	0	0	0	<Details algorithm="Support Vector M...	25	0	SINGLE	HUNG
4	CU12130	0.5505922406844544	0	0	0	<Details algorithm="Support Vector M...	25	0	SINGLE	DA...
5	CU5829	0.5487477467658215	0	0	0	<Details algorithm="Support Vector M...	25	0	SINGLE	DO...
6	CU1057	0.5475972506413769	1	42,562	5,000	<Details algorithm="Support Vector M...	25	1	WIDOWED	DIANE
33	CU9423	0.5374269846227704	1	25,923	350	<Details algorithm="Support Vector M...	62	1	DIVORCED	RIVA
34	CU6273	0.536681171179489	1	26,646	12,000	<Details algorithm="Support Vector M...	71	1	MARRIED	LAT...
35	CU9540	0.5366779461821751	0	0	0	<Details algorithm="Support Vector M...	25	0	SINGLE	FRE...
36	CU1564	0.5365665023147156	1	26,334	5,000	<Details algorithm="Support Vector M...	25	1	WIDOWED	CA...
37	CU6263	0.5365344995669501	1	37,618	5,000	<Details algorithm="Support Vector M...	25	1	OTHER	JIMMY
38	CU838	0.536468174245462	3	0	1,500	<Details algorithm="Support Vector M...	6,453	1	MARRIED	CH...
39	CU11047	0.536108785003169	2	33,212	1,300	<Details algorithm="Support Vector M...	31	1	DIVORCED	CARL
40	CU11183	0.5359801706423947	0	0	0	<Details algorithm="Support Vector M...	25	0	SINGLE	AMOS
41	CU15214	0.5359639335502011	3	0	200	<Details algorithm="Support Vector M...	18,981	1	MARRIED	MA

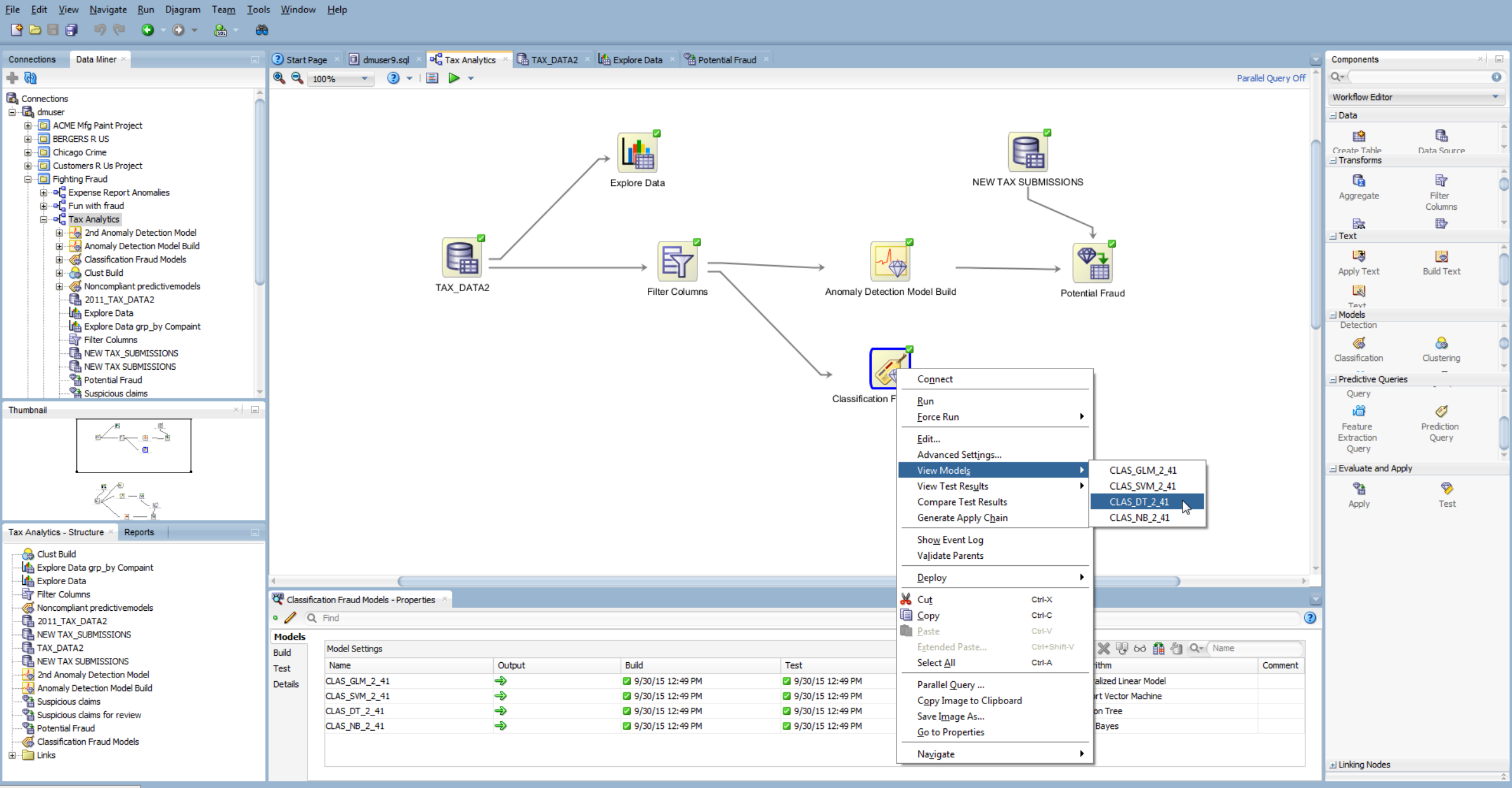
View Value

Find

```
<Details algorithm="Support Vector Machines" class="0">
   <Attribute name="MONEY_MONTLY_OVERDRAWN" actualValue="-123" weight=".077" rank="1"/>
   <Attribute name="MARITAL_STATUS" actualValue="WIDOWED" weight=".04" rank="2"/>
   <Attribute name="STATE" actualValue="NC" weight=".017" rank="3"/>
   <Attribute name="YRS_RESIDENCE" actualValue="4" weight=".014" rank="4"/>
   <Attribute name="SEX" actualValue="F" weight=".009" rank="5"/>
</Details>
```

Close

dmuser/Fun with Fraud/Tax Analytics

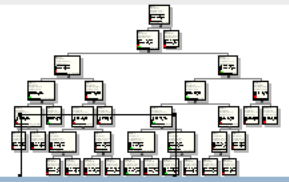


File Edit View Navigate Run Diagram Team Tools Window Help

Connections | Data Miner

- Connections
  - dmuser
    - ACME Mfg Paint Project
    - BERGERS R US
    - Chicago Crime
    - Customers R Us Project
    - Fighting Fraud
      - Expense Report Anomalies
      - Fun with fraud
      - Tax Analytics
        - 2nd Anomaly Detection Model
        - Anomaly Detection Model Build
        - Classification Fraud Models
        - Clust Build
        - Noncompliant predictivemodels
        - 2011_TAX_DATA2
        - Explore Data
        - Explore Data grp_by Compaint
        - Filter Columns
        - NEW TAX_SUBMISSIONS
        - NEW TAX SUBMISSIONS
        - Potential Fraud
        - Suspicious claims

Thumbnail



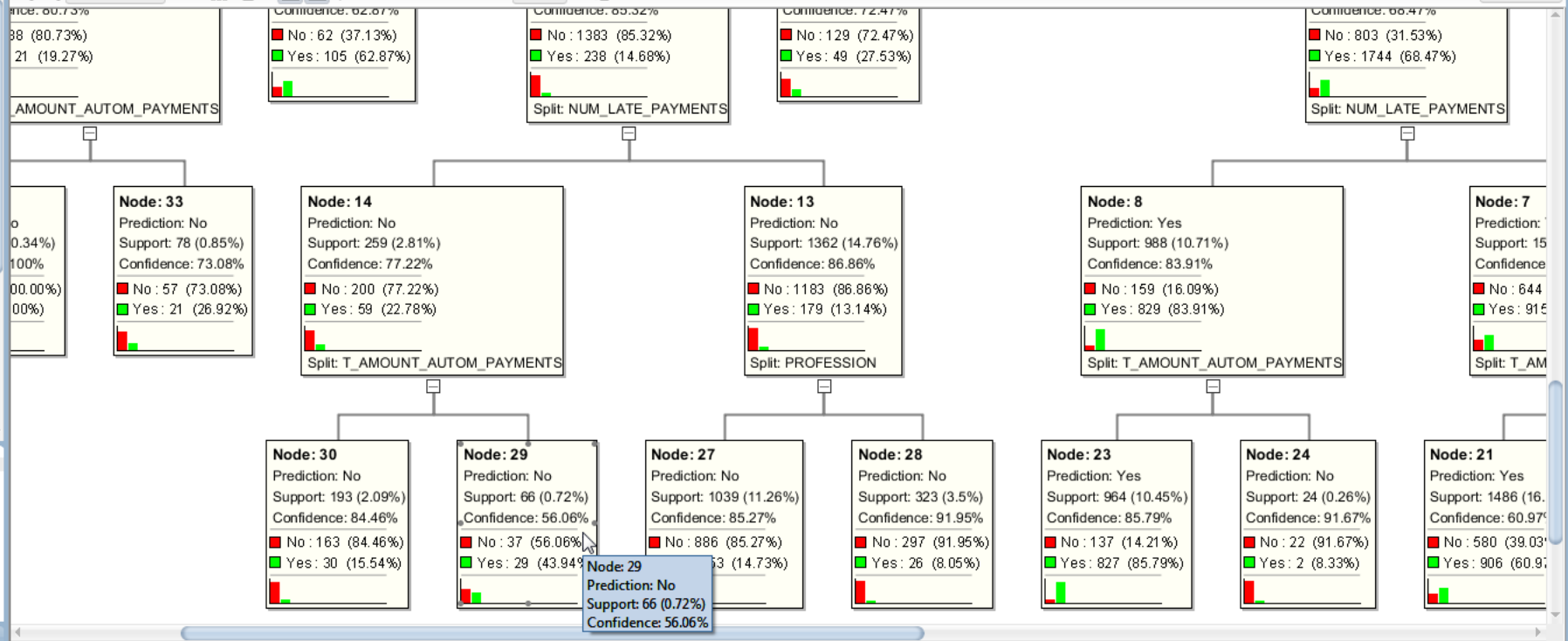
CLAS_DT_2_41 - Structure | Reports

- Node: 0
- Node: 17
- Node: 1
- Node: 2
- Node: 3
- Node: 4
- Node: 18
- Node: 5
- Node: 19
- Node: 20
- Node: 6
- Node: 7
- Node: 21
- Node: 22
- Node: 8
- Node: 23
- Node: 24

Start Page | dmuser9.sql | Tax Analytics | TAX_DATA2 | Explore Data | Potential Fraud | CLAS_DT_2_41

Tree | Settings

100% | Maximum Target Values: 2 | Save Rules...



Rule | Surrogates | Target Values

Node Rule: Wrap

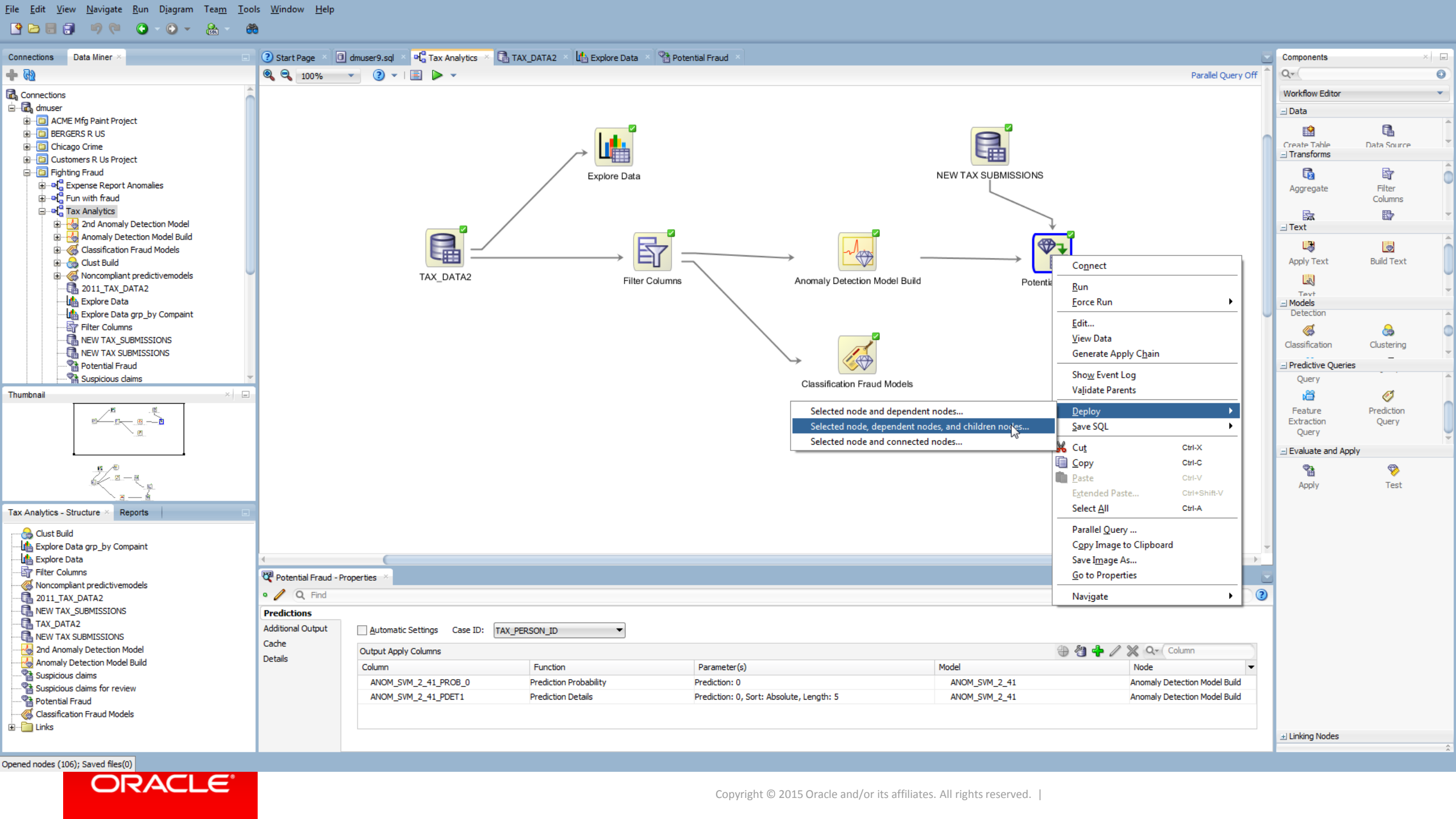
```
If   ITEMIZED_DEDUCTIONS > 246
And  ADJUSTED_DEDUCTIONS > 282
And  MONEY_MONTLY_OVERDRAWN <= 54.5
And  NUM_HEALTH_DEP <= 5.5
And  NUM_LATE_PAYMENTS > 4.5
And  T_AMOUNT_AUTOM_PAYMENTS <= 3800
Then No
```

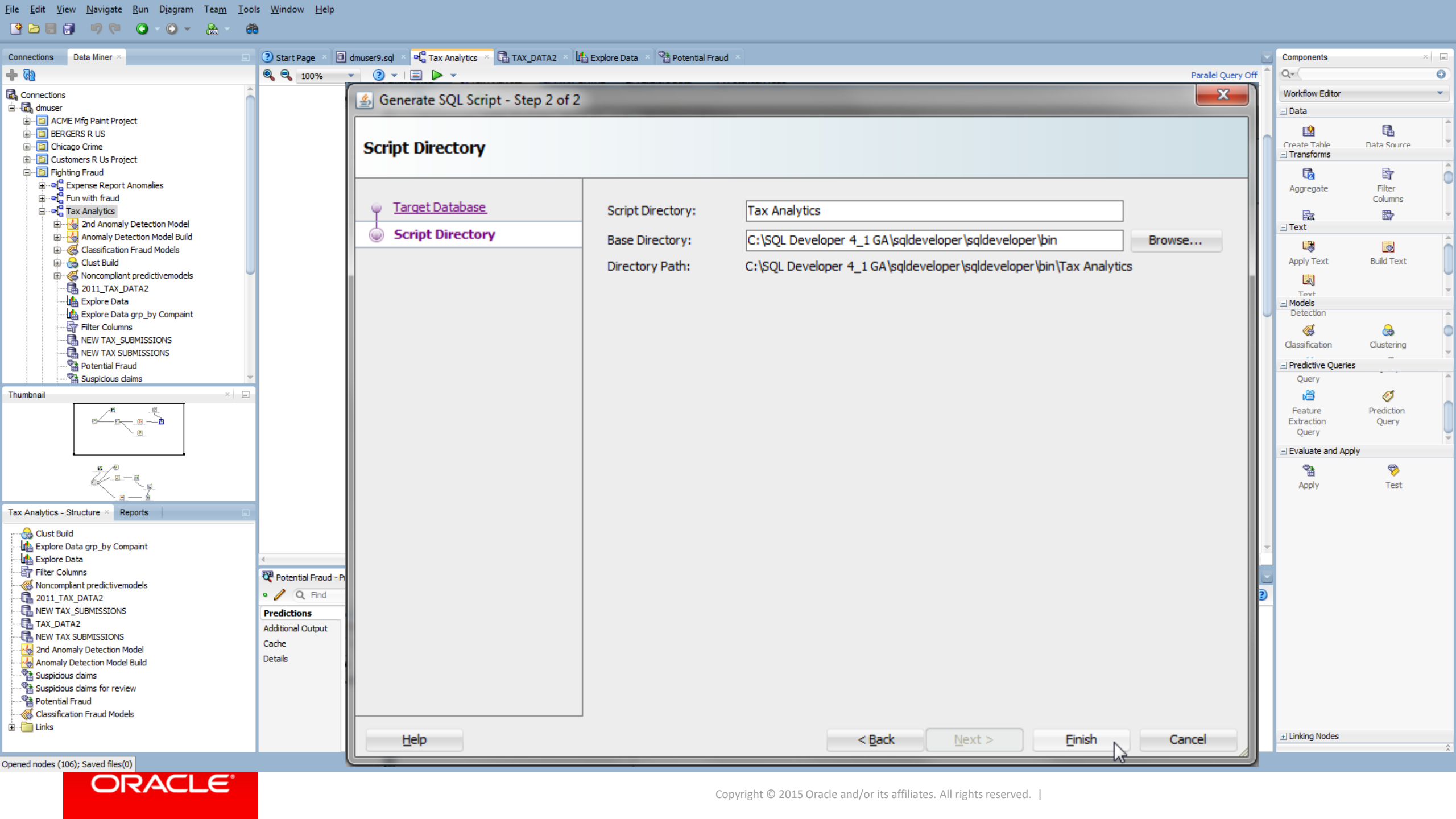
Confidence	0.5606060606060606
Support	0.007153696076306092

Components

No Components

Opened nodes (106); Saved files(0)





Fraud Prediction Demo

Automated In-DB Analytical Methodology

```
drop table CLAIMS_SET;
exec dbms_data_mining.drop_model('CLAIMSMODEL');
create table CLAIMS_SET (setting_name varchar2(30), setting_value varchar2(4000));
insert into CLAIMS_SET values ('ALGO_NAME','ALGO_SUPPORT_VECTOR_MACHINES');
insert into CLAIMS_SET values ('PREP_AUTO','ON');
commit;
```

```
begin
dbms_data_mining.create_model('CLAIMSMODEL', 'CLASSIFICATION',
  'CLAIMS', 'POLICYNUMBER', null, 'CLAIMS_SET');
end;
/
```

```
-- Top 5 most suspicious fraud policy holder claims
select * from
(select POLICYNUMBER, round(prob_fraud*100,2) percent_fraud,
      rank() over (order by prob_fraud desc) rnk from
(select POLICYNUMBER, prediction_probability(CLAIMSMODEL, '0' using *) prob_fraud
from CLAIMS
where PASTNUMBEROFCLAIMS in ('2to4', 'morethan4')))
where rnk <= 5
order by percent_fraud desc;
```

POLICYNUMBER	PERCENT_FRAUD	RNK
6532	64.78	1
2749	64.17	2
3440	63.22	3
654	63.1	4
12650	62.36	5

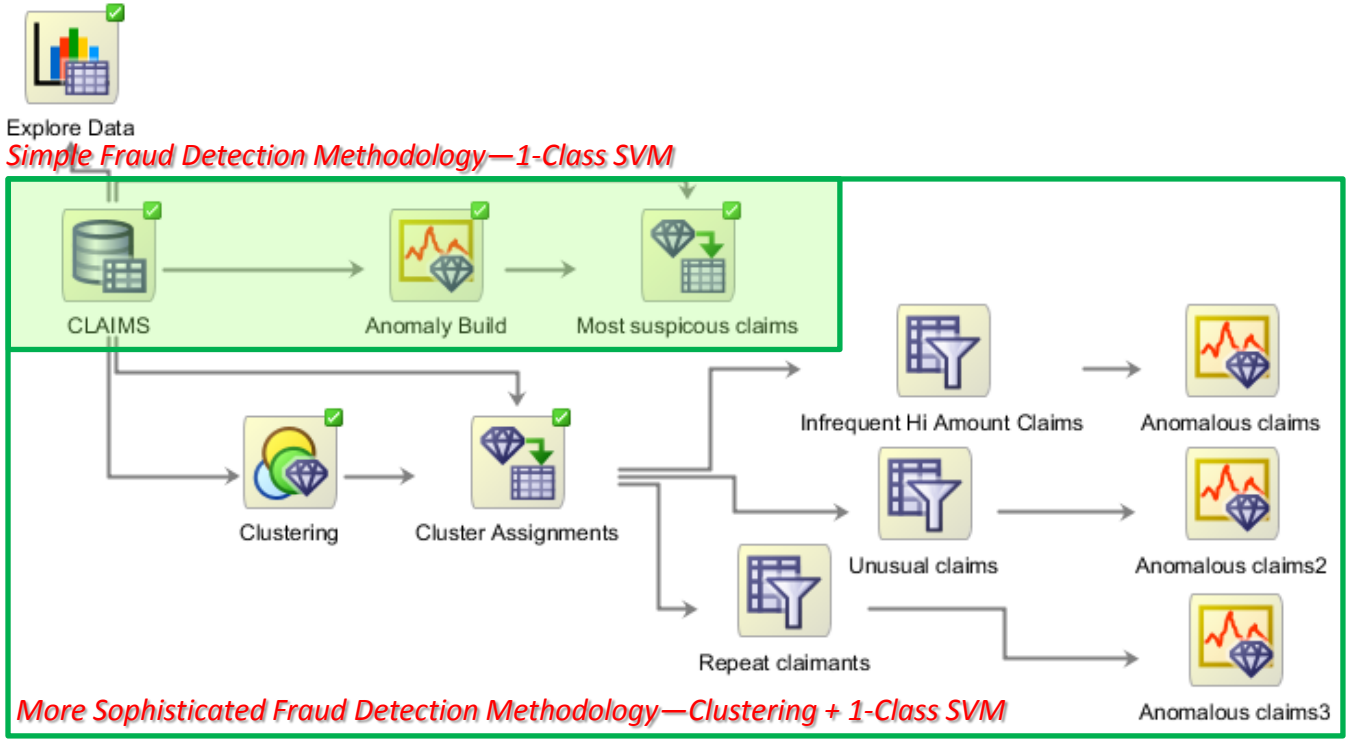
Automated Monthly "Application"! *Just add:*

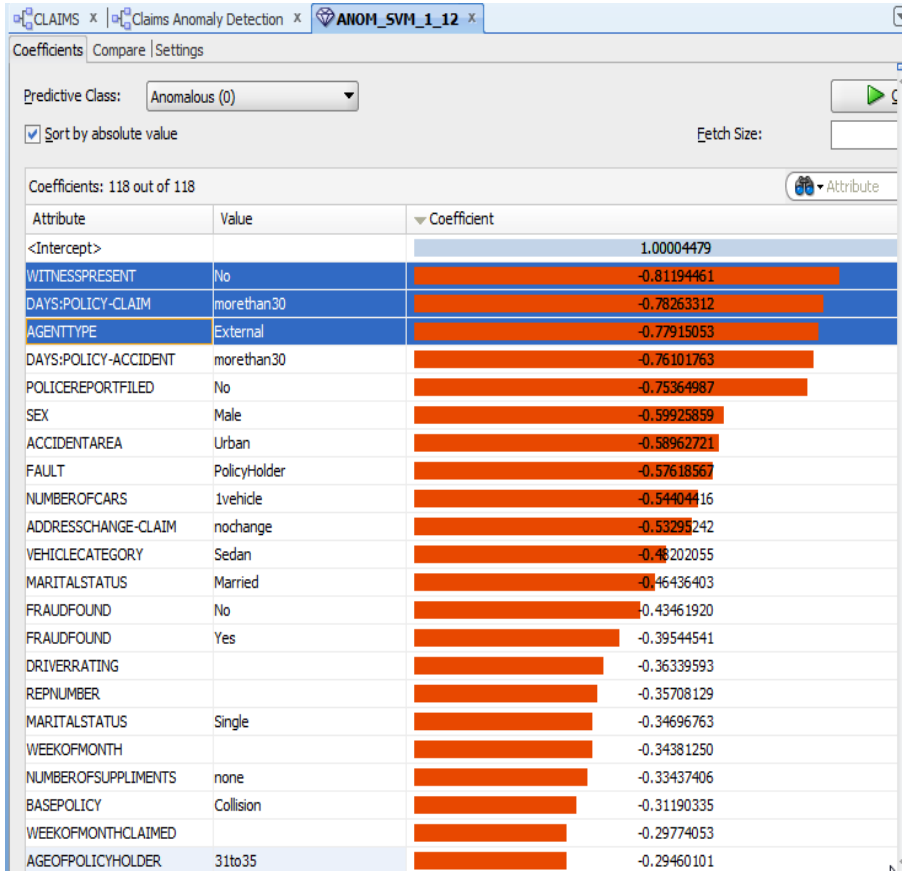
```
Create
View CLAIMS2_30
As
Select * from CLAIMS2
Where mydate > SYSDATE – 30
```

Time measure: set timing on;

Financial Sector/Accounting/Expenses

Anomaly Detection





Multiple Approaches To Detect Potential Fraud

1. Anomaly Detection (1-Class SVM)
   - Add feedback loop to purify the input training data over time and improve model performance
2. Classification
   - IF you have a lot of examples (25% or more) of fraud on which to train/learn
3. Clustering
   - Find records that don't high very high probability to fit any particular cluster and/or lie in the outlier/edges of the clusters
4. Hybrid of #3 and then #1
   - Pre-cluster the records to create "similar" segments and then apply anomaly detection models for each cluster
5. Panel of Experts
   - i.e. 3 out of 5 models predict possibly anomalous above 40% or any 1 out of N models considers this record unusual

ORACLE

Turkcell

Combating Communications Fraud

Objectives

- Prepaid card fraud—millions of dollars/year
- Extremely fast sifting through huge data volumes; with fraud, time is money

Solution

- Monitor 10 billion daily call-data records
- Leveraged SQL for the preparation—1 PB
- Due to the slow process of moving data, Turkcell IT builds and deploys models in-DB
- Oracle Advanced Analytics on Exadata for extreme speed. Analysts can detect fraud patterns almost immediately

- "Turkcell manages 100 terabytes of compressed data—or one petabyte of uncompressed raw data—on Oracle Exadata. With Oracle Data Mining, a component of the Oracle Advanced Analytics Option, we can analyze large volumes of customer data and call-data records easier and faster than with any other tool and rapidly detect and combat fraudulent phone use."
  – Hasan Tonguç Yılmaz, Manager, Turkcell İletişim Hizmetleri A.Ş.

Oracle Advanced Analytics In-Database Fraud Models

Exadata

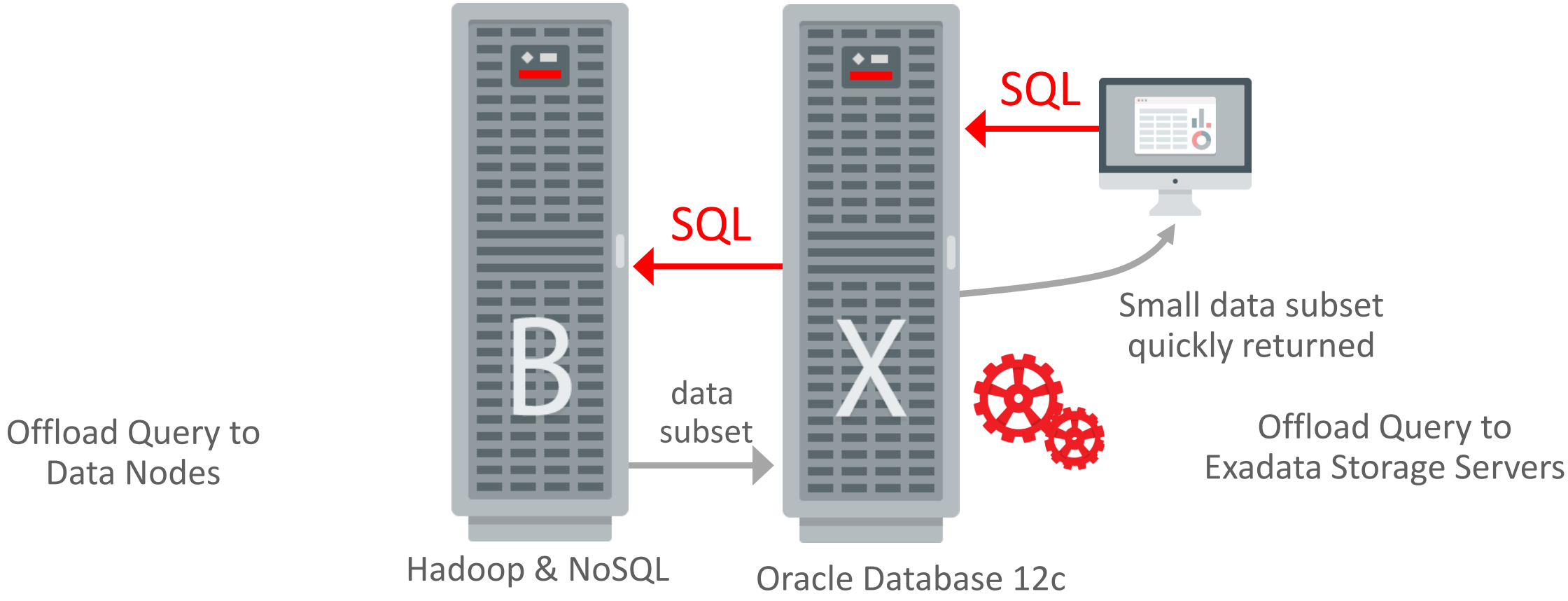
Big Data SQL

Push down SQL predicts to storage layers

ORACLE

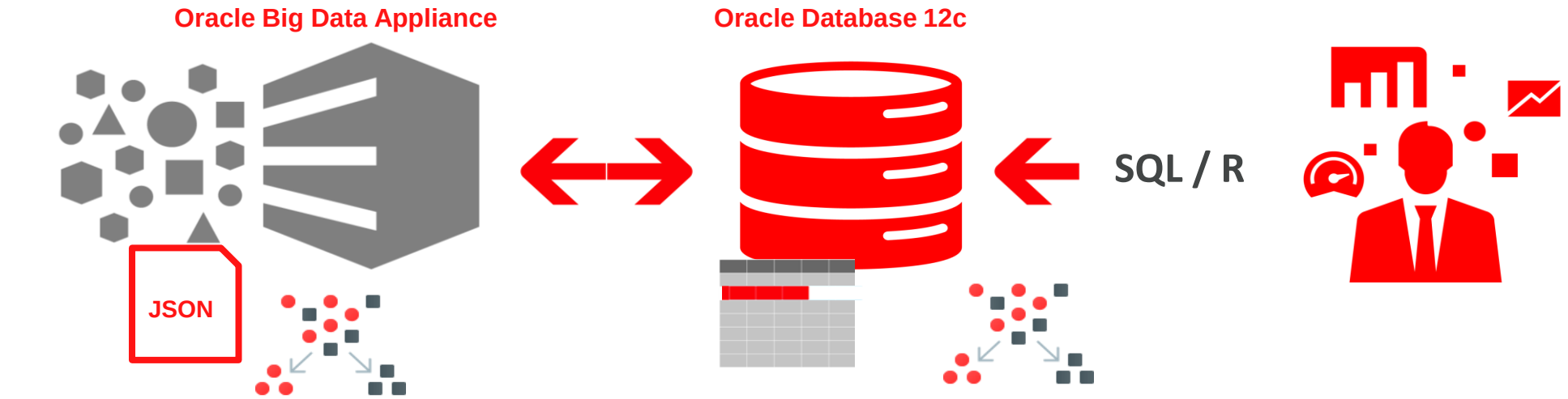
Introducing Oracle Big Data SQL

Massively Parallel SQL Query across Oracle, Hadoop and NoSQL



Manage and Analyze All Data—SQL & Oracle Big Data SQL

Oracle's Advanced Analytics



Structured and Unstructured Data Reservoir
- JSON data
- HDFS / Hive
- NoSQL
- Spatial and Graph data
- Image and Video data
- Social Media

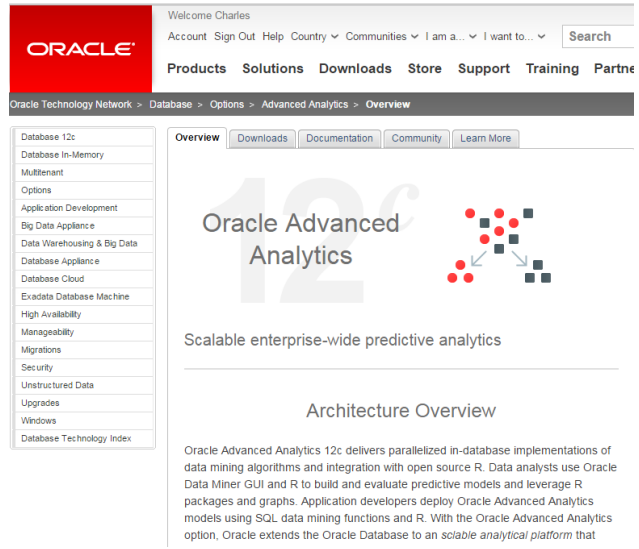
Store business-critical data in Oracle
- Customer data
- Transactional data
- Unstructured documents, comments
- Spatial and Graph data
- Image and Video data
- Social Media

Data analyzed via SQL / R / GUI
- R Clients
- SQL Clients
- Oracle Data Miner

Getting started

OAA Links and Resources

- **Oracle Advanced Analytics Overview:**
  - **OAA presentation—** Big Data Analytics in Oracle Database 12c With **Oracle Advanced Analytics** & Big Data SQL
  - **Big Data Analytics with Oracle Advanced Analytics: Making Big Data and Analytics Simple white paper** on OTN
  - *Oracle Internal* OAA Product Management Wiki and Workspace
- ***YouTube* recorded OAA Presentations and Demos:**
  - **Oracle Advanced Analytics and Data Mining at the YouTube Movies**
    (6 + OAA "live" Demos on ODM'r 4.0 New Features, Retail, Fraud, Loyalty, Overview, etc.)
- **Getting Started:**
  - Link to **Getting Started w/ ODM blog entry**
  - Link to New OAA/Oracle Data Mining 2-Day Instructor Led Oracle University course.
  - Link to **OAA/Oracle Data Mining 4.0 Oracle by Examples (free) Tutorials** on OTN
  - Take a Free Test Drive of Oracle Advanced Analytics (Oracle Data Miner GUI) on the Amazon Cloud
  - Link to **OAA/Oracle R Enterprise (free) Tutorial Series** on OTN
- **Additional Resources:**
  - Oracle Advanced Analytics Option on OTN page
  - **OAA/Oracle Data Mining on OTN** page, ODM Documentation & ODM Blog
  - **OAA/Oracle R Enterprise page on OTN** page, ORE Documentation & ORE Blog
  - Oracle SQL based Basic Statistical functions on OTN
  - BIWA Summit'16, Jan 26-28, 2016 – Oracle Big Data & Analytics User Conference @ Oracle HQ Conference Center



BIWA SUMMIT 2016

The Oracle Big Data + Analytics User Conference
January 26-28, 2016
Including Oracle Spatial Summit

Home | Abstract Submission | Sponsorship | Hotel and Travel | Registration Pricing | Registration

BIWA SIG
Better Information Better Results
www.oraclebiwa.org

ORACLE SPATIAL AND GRAPH SPECIAL INTEREST GROUP (SIG)

NoCOUG — Northern California Oracle Users Group

Publicity

- Oracle Business Analytics Newsletter
- DB Insider Dec 2014
- Oracle Magazine
- Latest BIWA SIG Blog Entry
- Jeff Shauer Blog Entry
- Daily BIWA Newsletter
- Email to BIWA members
- Real Time BI Webcast
- Oracle Events Calendar
- Oracle ACE Newsletter
- DB Insider Jan 2015 with Spatial Summit
- Lots of other emails

January 26-28, 2016

Oracle Conference Center at Oracle HQ Campus, Redwood Shores, CA

- **Hands-on-Labs**
- **Customer stories, told by the customers**
- **Educational sessions by Practitioners and Direct from Developers**
- **Oracle Keynote presentations**
- **Presentations covering: Advanced Analytics, Big Data, Business Intelligence, Cloud, Data Warehousing and Integration, Spatial and Graph, SQL**
- **Networking with product management and development professionals**

ORACLE


ORACLE®