Use Case — Retail
Intel Big Data Enablement — Marketing
White Goods Prediction

Intel and Cloudera Help a Company Use Predictive Analytics to Foresee Future Purchasing Behaviors

Intel and Cloudera implement a solution to ingest multiple data sources, using the information to create variables to predict buying trends among customers.

Why Intel and Cloudera

Intel and Cloudera take the guesswork out of Hadoop. Using a unique collaborative approach, we deliver excellent performance, security, and quality distribution, built on open standards. Working with more vendors across the ecosystem, only a solution built on CDH can ensure freedom from lock-in, enabling you to build a robust big data solution to meet the needs of your business today and into the future.

- Uniquely aligned product roadmaps for software and hardware to drive innovation faster, providing more industry firsts than any other Hadoop alternative.
- Deep partnerships with virtually every provider in the data center, streamlining the process for building Big Data solutions.
- Proven track records of identifying the driving industry standards, so you don't run the risk of stranding yourself on an island.

A French company that sells household electronic appliances and white goods seeks a solution to predict the product-buying habits of previous customers, identifying the likelihood of households to make additional purchases.

The Company deploys Cloudera distribution of Hadoop (CDH) to ingest and process large amounts of raw customer data, such as previous purchase history, demographics, service requests, and more.

The new Cloudera-based solution creates over a hundred deterministic variables from the customer data, using complex random forest model algorithms. CDH harnesses these variables in a Bayesian process that transforms them into a usable solution to accurately predict buying trends among customers, empowering the Company to make educated, data-driven business decisions.

Results

The Intel/Cloudera solution yields the following benefits:

- Provides the company with a method of creating variables from customer data, which can be used to predict the purchase patterns of each customer.
- Uses these variables to accurately hindcast customers' purchasing trends for a previous historical year, resulting in consistent predictions when compared to these customers' actual purchases.
- Predicts which households have the highest likelihood of purchasing a product, for each household segment.
- For each of the top selected households, determines which product groups customers are most likely to purchase from.

Business drivers

Executives from the Company approached Intel with the primary goal of answering the following questions:

- How can we identify the top 500 customer households that are most likely to purchase appliances within the next 12 months?
- From which product lines are the top households most likely to make a purchase?

Answers to these questions would help the Company plan better and construct a business strategy for the purchasing trends of their customers.

Solution details

The Company contacted Intel to help solve its Big Data needs. Based on Intel's recommendations, they decided to use a Hadoop cluster running CDH and R statistical software (*Figure 1*).

The CDH cluster loads and wrangles data from nine different sources. The files from these data sources—once merged and aggregated—were used to create 106 variables, which were defined to cover the following:

- Basic customer information
- Household information
- Household segmentation
- Previous purchase behavior
- Service and maintenance request behavior
- Household accommodations
- Government records of household and local demographics
- Events triggered by service requests
- Resident activity

To determine and select which variables had the strongest impact on customers' purchasing habits, we applied a random forest algorithm for each segment of customer data separately. The random forest algorithm contains 1,000 different trees.

The theory behind the analytics is based on Bayesian inference, a statistical method of inference to calculate the posterior probability of a customer's interest.

Each variable's importance is determined by the "mean decrease accuracy" value given to it by the results of the random forest algorithm.

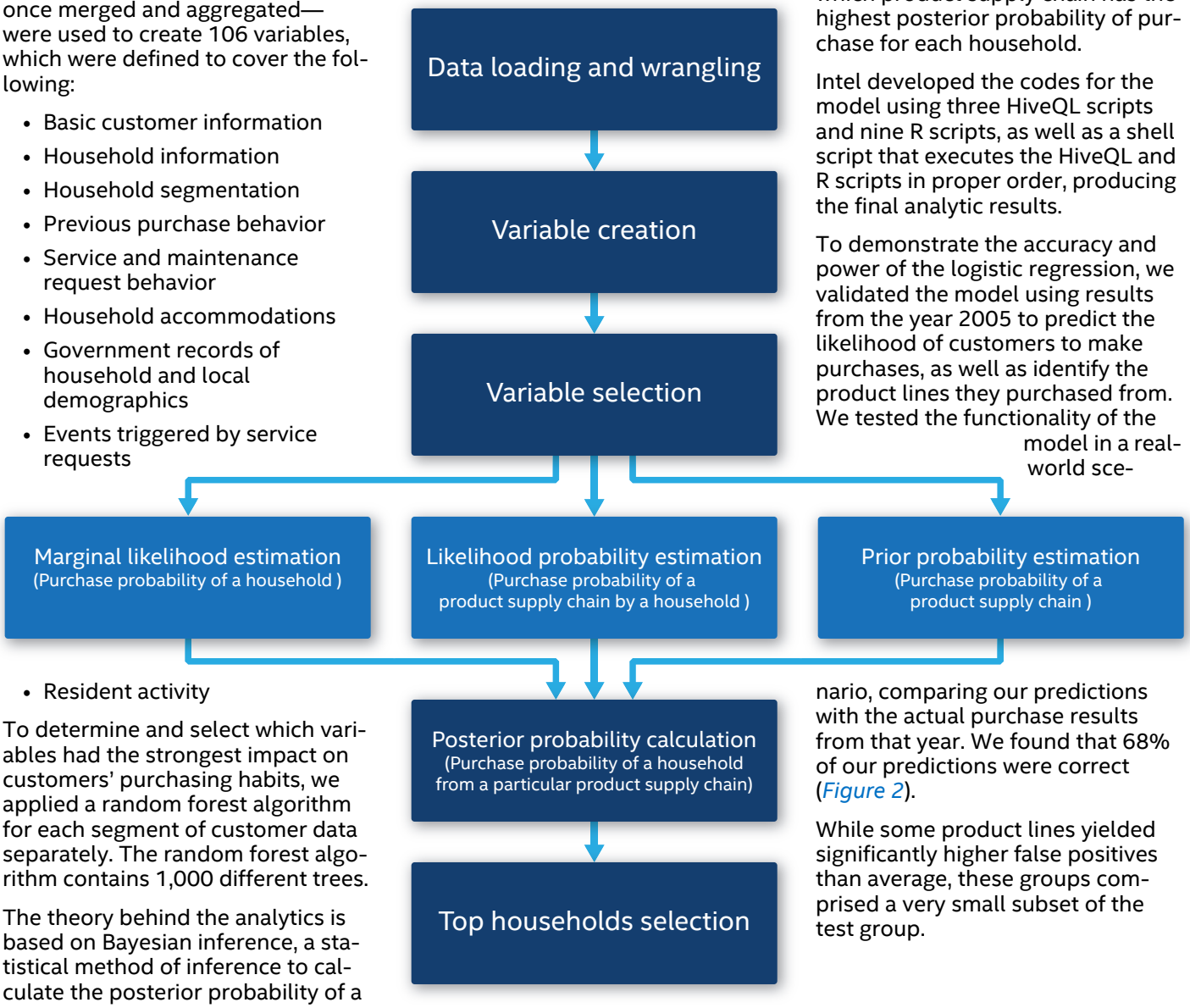


Figure 1 **Pruning the random forest.** After wrangling data from nine sources, the solution creates 106 variables to determine the probabilities of customers/households to make additional purchases from various product lines. The model executes Hive queries, converts Hive tables into CSV files, executes R scripts, and saves the results as CSV files. From there, the model sorts selections hierarchically based on probability likelihood scores.

The logistic regression model is built upon the variables by priority of importance, as determined by the random forest algorithm. Each customer has a "buy/no-buy" flag that indicates either a positive or negative impact on the likelihood of a household to make a purchase. It also can help determine—very accurately—the product line from which a customer will buy.

Households are ordered based on their probability of purchase, from highest to lowest, as calculated by the model. The model also predicts which product supply chain has the highest posterior probability of purchase for each household.

Intel developed the codes for the model using three HiveQL scripts and nine R scripts, as well as a shell script that executes the HiveQL and R scripts in proper order, producing the final analytic results.

To demonstrate the accuracy and power of the logistic regression, we validated the model using results from the year 2005 to predict the likelihood of customers to make purchases, as well as identify the product lines they purchased from. We tested the functionality of the model in a real-world scenario, comparing our predictions with the actual purchase results from that year. We found that 68% of our predictions were correct (*Figure 2*).

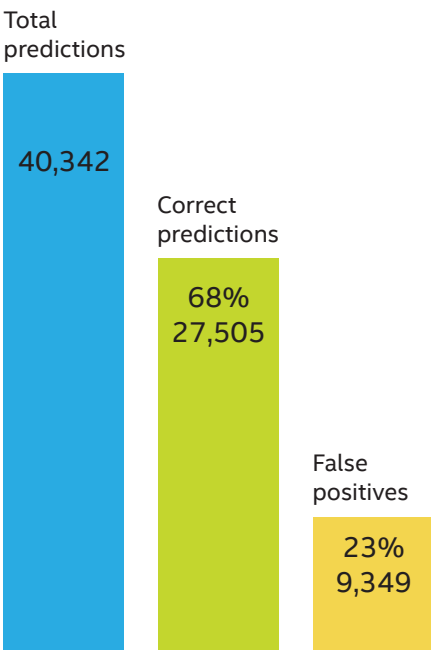
While some product lines yielded significantly higher false positives than average, these groups comprised a very small subset of the test group.

Cloudera Enterprise

Cloudera provides a secure and fault-tolerant platform to accurately predict which customers will make purchases, and from what product line, enabling the company to make better, smarter, faster, data-driven business decisions.

Data scientists can build sound predictive analytical models through data mining and produce business intelligence solutions running on CDH.

Figure 2 **Hindcasting the future.** Based on hindcasts using historical and current data, the Company has been able to correctly predict customer purchases 68% of the time.



Hadoop allows the data scientist access to tools such as Mahout, a scalable machine-learning library, as well as streaming, which allows creating and running MapReduce jobs using Python or another executable script.

Cloudera distribution of Hadoop provides an efficient and cost-effective platform for Big Data solutions.

Summary

Before deploying CDH, the Company's predictive methods were unable to provide the answers they sought. With Intel/Cloudera, the Company now reaps the benefits of an advanced and highly consistent predictive model.

The regression model Intel helped develop for the Company hindcasts purchases for a historical year to validate model function accuracy. It has already demonstrated its proficiency. With this process in place, the Company can now accurately identify which households are most likely to purchase white goods within the next 12-month period, and which product lines those white goods will be selected from. The Company is also now able to see at a glance the number of purchases per family and the average time between purchases. This clearer visibility of customers and their purchasing habits has helped the Company plan better.

Let us help your business too.

Spotlight on Cloudera

Cloudera is revolutionizing enterprise data management by offering the first unified Platform for Big Data, an enterprise data hub built on Apache Hadoop™. Cloudera offers enterprises one place to store, access, process, secure, and analyze all their data, empowering them to extend the value of existing investments while enabling fundamental new ways to derive value from their data.

Cloudera's open source Big Data platform is the most widely adopted in the world, and Cloudera is the most prolific contributor to the open source Hadoop ecosystem. As the leading educator of Hadoop professionals, Cloudera has trained over 40,000 individuals worldwide. Over 1,600 partners and a seasoned professional services team help deliver greater time to value. Finally, only Cloudera provides proactive and predictive support to run an enterprise data hub with confidence. Leading organizations in every industry plus top public sector organizations globally run Cloudera in production.

For more information, visit **www.cloudera.com**.

cloudera®

Meeting your needs

We look forward to meeting with you to define your requirements and meet your objectives.

- **Accelerate time to value:** Achieve real-time cost savings, respond to market trends, and drive innovation.
- **Secure Big Data:** Deploy a sustainable Big Data program that doesn't put your organization, or you, at risk.
- **Maintain control:** Work with a partner who educates your team so you become self-sufficient.
- **Increase business potential**: Create and execute a plan that helps you adapt now, and in the future.

Hadoop sizing guide

		Cluster size		
		Small	**Medium**	**Large**
CPU		**Intel® Xeon® Processor E5 v3**		
Storage (TB)		<72 TB	72 to 570 TB	>570 TB
Node count	**Master**	2 to 3	4 to 7	≥8
	Slaves	<12	12 to 95	≥ 96
Memory (GB)	**Master**	64 GB	128 GB	≥256 GB
	Slaves	48 GB	96 GB	≥128 GB
Network		1 Gbps	10 Gbps	10 Gbps

Hardware configuration is highly dependent on workload. A high storage density cluster may be configured with a 4 TB JBOD hard disk, while a compute intensive cluster may be configured with a higher memory configuration.

Contact us

Contact your sales rep or e-mail us.

Intel.com/bigdata/services

cloudera®

intel®

332921-001