Perbandingan Analisis Faktor Penentu Penjualan PT. X Menggunakan *LASSO Regression* dan *Gradient Boosted Regression Tree*

Jessica Athalia, Henry Novianus Palit, Silvia Rostianingsih
Program Studi Informatika, Fakultas Teknologi Industri, Universitas Kristen Petra
Jl. Siwalankerto 121 – 131 Surabaya 60236
Telp. (031) – 2983455, Fax. (031) – 8417658
E-Mail: jessica.athalia22@gmail.com, hnpalit@petra.ac.id, silvia@petra.ac.id

ABSTRAK

Informasi menjadi aset yang sangat penting dalam sebuah organisasi. Namun, staf perusahaan PT. X merasa kesulitan untuk menganalisis data karena harus diolah satu per satu. Terlebih, proses analisis data yang dilakukan langsung di database operasional dapat memperlambat kinerja database operasional tersebut. Kemudian, ketika *Board of Director* ingin mengetahui alasan dibalik performa penjualan, para pegawai menyimpulkan faktor-faktor penyebabnya berdasarkan asumsi belaka. Penelitian ini mengimplementasikan data warehouse dengan bantuan ETL *tools* kemudian melakukan analisis terhadap data transaksi penjualan PT. X untuk mengetahui faktor-faktor yang memiliki pengaruh dalam menentukan tingkat penjualan produk-produknya. Model faktor dibentuk untuk brand-brand perusahaan PT. X yang tingkat penjualannya tidak memuaskan selama beberapa tahun belakangan. Faktor yang diuji antara lain harga produk, ketersediaan stok barang, ketepatan waktu pengiriman barang, jumlah barang retur, bulan pemesanan, dan harga bahan baku. Analisis faktor penentu ini dilakukan menggunakan 2 metode yaitu, *LASSO regression* dan *Gradient Boosted Regression Tree*. Hasil analisis dari kedua metode diuji berdasarkan nilai *Root Mean Squared Error*, *R-squared*, dan *Variance Inflation Factor* untuk mengetahui model faktor yang lebih baik. Berdasarkan pengujian, analisis menggunakan metode *LASSO regression* dan *Gradient Boosted Regression Tree* berhasil memilih faktor-faktor yang berpengaruh signifikan terhadap penjualan PT. X. Akan tetapi, model faktor dari *Gradient Boosted Regression Tree* menunjukkan tingkat akurasi yang lebih baik daripada *LASSO regression*. Supaya perusahaan dapat melakukan analisis faktor penentu di kemudian hari, sebuah program dibuat dengan metode *Gradient Boosted Regression Tree*.

Kata Kunci: Data warehouse; ETL; *LASSO regression*; *Gradient Boosted Regression Tree*; data transaksi penjualan

ABSTRACT

Information becomes a crucial asset for an organization. However, employees of PT. X are facing difficulty in analyzing data because it has to be processed one by one. Moreover, analyzing data in an operational database is not recommended as it can interfere with the performance of the operational database. Then, when the Board of Directors want to know the reason behind its sales' performance, they conclude it based on their mere assumption. This research implemented a data warehouse with the help of ETL tools. Then, sales transactions of PT. X were analyzed to get information about factors that affect company's revenue. Factor models were formed for brands which sales were not good enough these past few years. Factors which are examined are sales price, stock availability, on time delivery of goods, quantity of returns, month of transaction, and cost price. The analysis was carried with two methods, LASSO regression and Gradient Boosted Regression Tree. These models were measured by Root Mean Squared Error, R-squared, and Variance Inflation Factor to know which model performs better. Result of the research shows LASSO regression and Gradient Boosted Regression Tree succeed in performing feature selection for sales transactions of PT. X. Yet, the factor model from Gradient Boosted Regression Tree gives a better result than LASSO regression. Last, a program was made for the company in the need of future analysis using Gradient Boosted Regression Tree.

Keywords: *Data warehouse; ETL; LASSO regression; Gradient Boosted Regression Tree; sales transactions*

1. PENDAHULUAN

Informasi menjadi aset yang sangat penting dalam sebuah organisasi. Namun, organisasi saat ini dihadapkan pada volume data yang sangat masif. Melakukan *analytical query* langsung di database operasional dapat memperlambat kinerja database operasional tersebut. Oleh karenanya, dibutuhkan data warehouse. Data warehouse menyediakan data yang terintegrasi dan konsisten untuk mendukung proses pengambilan keputusan. Fokus dari data warehousing adalah mengintegrasikan berbagai sumber data sehingga proses analisis data dapat dilakukan guna memperoleh informasi yang dibutuhkan tanpa mengganggu kerja dari database operasional [6].

PT. X adalah perusahaan yang memproduksi lem dan cat. PT. X didirikan pada tahun 1976 dan memiliki kantor pusat di Surabaya. Sejak tahun 2013, PT. X mendirikan pabrik di Jakarta untuk melayani pasar ASEAN. PT. X mulai menggunakan SQL Server sejak tahun 2015. Setiap *salesperson* difasilitasi dengan *mobile apps* sejak tahun 2017 untuk mencatat data *take order* di MySQL kemudian data akan masuk ke SQL Server jika telah divalidasi oleh bagian *finance* dan menjadi *sales order*. Adapun data *Purchase Order*, *Inventory*, dll diinputkan secara manual oleh pegawai di masing-masing bagiannya ke SQL Server. Selain data tersebut, PT.X juga masih menggunakan excel untuk mencatat data transaksi promosi. Itu menyebabkan hasil dari promosi yang dilakukan tidak dapat terlihat dalam analisis penjualan.

Terlebih lagi, kemungkinan besar data perusahaan bisa berbentuk file lainnya di masa yang akan datang sehingga menimbulkan

concern dari perusahaan. Data yang harus diolah satu-satu membuat para staf merasa kesulitan untuk bisa menganalisis data guna mengetahui apa yang terjadi pada suatu proses bisnis. Dari sana muncul kesadaran perusahaan akan kebutuhan untuk membangun data warehouse. Lalu, ketika *Board of Director* ingin mengetahui mengapa hal tersebut terjadi, para pegawai menyimpulkan faktor-faktor penyebabnya berdasarkan asumsi belaka. Padahal, pemimpin membutuhkan laporan tersebut supaya bisa mengambil keputusan strategis yang tepat.

Masalah tersebut akan diselesaikan dengan membangun data warehouse menggunakan *ETL tools* dan mengetahui faktor-faktor yang berpengaruh dalam menentukan tingkat penjualan suatu produk menggunakan analisis regresi terhadap data yang ada sehingga perusahaan dapat mengambil langkah yang tepat untuk meningkatkan performa penjualan produk-produknya. Penelitian terkait menentukan faktor-faktor yang berpengaruh terhadap suatu variabel menggunakan analisis regresi pernah dilakukan sebelumnya oleh Germa Bel dan Mildred [1] yang menguji hipotesis faktor yang mempengaruhi terjadinya *intermunicipal cooperation* menggunakan *meta-regression analysis*. Penelitian lain [2] melakukan integrasi data dengan SQL Server Integration Services (SSIS) kemudian menggunakan metode Naïve Bayes untuk mengetahui berbagai faktor/*event* saat menyetir yang berpengaruh besar dalam menentukan rata-rata konsumsi bahan bakar. Penelitian serupa [3] juga dilakukan yaitu menentukan faktor-faktor yang mempengaruhi konsumsi bahan bakar pada alat transportasi umum menggunakan metode *multiple regression*.

Penelitian ini akan menganalisis pengaruh suatu faktor terhadap penjualan suatu produk menggunakan metode *LASSO regression* dan *Gradient Boosted Regression Tree* (GBRT) hingga terbentuk model faktor yang lebih akurat. *LASSO regression* dipilih karena merupakan metode regresi khusus yang bisa melakukan *feature selection* sehingga bisa mendiskualifikasi faktor yang tidak relevan secara otomatis. Sementara, *Gradient Boosted Regression Tree* merupakan metode regresi lainnya yang bersifat fleksibel karena mewarisi sifat dari *regression tree* yang *high interpretability*, *conceptual simplicity*, dan *computational efficiency* dengan *boosting approach* yang bisa meningkatkan keakuratan model faktor [11].

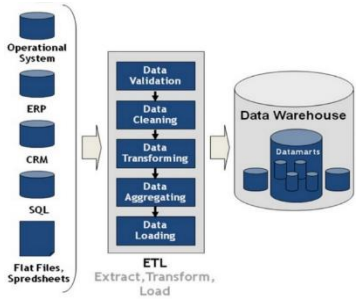
2. DASAR TEORI

2.1 Data Warehouse

Data warehouse adalah sekumpulan data dalam jumlah besar dari berbagai macam sumber yang disimpan dalam 1 tempat penyimpanan data. Data warehouse menyimpan *historical data* yang dapat dianalisis untuk membantu suatu perusahaan dalam membuat suatu keputusan. Ada 2 pandangan berbeda dari 2 orang pionir data warehouse yaitu, Bill Inmon dan Ralph Kimball, tentang cara membangun data warehouse. Ralph Kimball memiliki konsep *bottom-up approach* sementara konsep data warehouse menurut Bill Inmon adalah *top-down approach* [9].

Top-down approach menampilkan data keseluruhan bisnis sebagai satu kesatuan dan kemungkinan redundansi data rendah namun membutuhkan waktu yang lama dan tim professional yang lebih besar untuk membangun data warehouse karena tingginya kompleksitas model data. Sedangkan, *bottom-up approach* lebih sederhana karena berfokus pada proses bisnis bukan *whole enterprise* sehingga membutuhkan waktu lebih cepat untuk membangun data warehouse meskipun redundansi data menjadi relatif lebih tinggi karena data didenormalisasi [9]. Sebelum mengimplementasikan data warehouse, model data perlu dirancang terlebih dahulu sebagai panduan arsitektur data warehouse yang akan dibangun nantinya. Ada 3 model data yang paling populer untuk membentuk data warehouse antara lain, *star schema*, *snowflake schema*, dan *galaxy schema* [12].

Proses utama dalam pembuatan data warehouse adalah proses *extract*, *transform*, dan *load* (ETL) seperti Gambar 1. *Extract* adalah proses memilih dan mengambil data dari satu atau beberapa sumber. Dalam proses *extract*, data akan divalidasi untuk memeriksa kebenaran dari nilai data yang diperoleh. *Transform* adalah proses membersihkan, mengubah data agar data bersifat lengkap dan konsisten serta menggabungkan data yang didapatkan dari proses *extract*. Jadi, proses *transform* juga memperbaiki data seperti melakukan standardisasi dataset, mengisi *field* yang masih kosong, dll. *Load* adalah proses yang berfungsi untuk memasukkan data ke dalam database target yang disebut data warehouse. Ada 2 cara untuk melakukan *load* data ke dalam data warehouse yaitu, *full load* dan *incremental load*. *Full load* adalah memasukkan seluruh data dalam satu waktu seperti saat pertama kali. *Incremental load* adalah memasukkan data secara rutin dalam jangka waktu tertentu dan data yang di *load* hanya data baru/perubahan data yang terjadi setelah tanggal terakhir kali melakukan *load* data [13].



Gambar 1. Proses ETL [4]

2.2 *LASSO (Least Absolute Shrinkage and Selection Operator) Regression*

Multiple linear regression digunakan untuk menganalisis *influence degree* dari berbagai macam faktor. Namun, masalah *multicollinearity* sering muncul sehingga mempengaruhi *fitting accuracy* dari model yang terbentuk [14]. Untuk menangani masalah *multicollinearity* yang sering terjadi, *LASSO regression* diaplikasikan dalam menentukan koefisien regresi dari *multiple linear regression* menggunakan persamaan berikut.

$$\hat{\beta}_{LASSO} = \arg\min_{\beta}\left\{ |y - X\beta|^2 + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \right\} \quad with \ \lambda \geq 0 \tag{1}$$

λ adalah parameter yang mengontrol seberapa banyak datanya disusutkan. Semakin tinggi nilai λ, semakin besar penyusutan variabel-variabelnya. *LASSO regression* memungkinkan koefisien dari variabel yang tidak relevan bisa mencapai 0. Hal itu membuat *LASSO regression* bisa melakukan *variable selection* secara otomatis. Ada 2 metode yang biasa digunakan untuk menentukan nilai λ yang tepat, yaitu *cross-validation* dan *information criteria*. *Cross validation* adalah metode yang paling sering digunakan untuk menentukan λ dari *LASSO regression* [8]. *Cross-validation* membagi data menjadi k bagian kemudian menghitung *mean squared error*. Nilai λ dengan *mean squared error* terendah akan dipilih [7].

2.3 *Gradient Boosted Regression Tree*

Gradient Boosted Regression Tree merupakan metode *gradient descent* yang dimodifikasi dengan *boosting algorithm* untuk meningkatkan akurasi dari model yang terbentuk [5].

$$F_m(x) = F_{m-1}(x) + vh_m(x) \quad v \in [0;1] \qquad (2)$$

Parameter *shrinkage* v mengontrol *learning rate* yaitu berapa banyak prosedur ini dilakukan. Parameter v berhubungan erat dengan iterasi *boosting*. Semakin kecil nilai v menandakan semakin banyak iterasi *boosting* dilakukan sampai menemukan *optimal fit* dengan *error rate* terendah [10].

3. DESAIN SISTEM

3.1 Analisis Perusahaan

Data perusahaan disimpan dalam bentuk tabel dan penjelasan mengenai tabel-tabel yang diberikan terdapat pada Tabel 1.

Tabel 1. Keterangan Tabel Data Perusahaan

Nama Tabel	Keterangan
PURCHTABLE	Menyimpan data master dari pembelian
PURCHLINE	Menyimpan detail transaksi pembelian
SALESTABLE	Menyimpan data master dari penjualan
SALESLINE	Menyimpan detail transaksi penjualan
CUSTPACKING SLIPJOUR	Menyimpan data master dari barang yang sudah dikirim
CUSTPACKING SLIPTRANS	Menyimpan detail dari barang yang sudah dikirim
CUSTINVOICE JOUR	Menyimpan data master dari nota barang terkirim
CUSTINVOICE TRANS	Menyimpan detail transaksi dari nota barang terkirim
ECORES PRODUCT	Menyimpan data master item
INVENTTABLE	Menyimpan detail item
INVENTSUM	Menyimpan status stok barang
AF_BRAND ITEMS	Menyimpan daftar brand produk perusahaan
AF_GOLONGAN ITEMS	Menyimpan daftar golongan produk perusahaan
AF_JENISITEMS	Menyimpan daftar jenis produk perusahaan
AF_KATEGORI ITEMS	Menyimpan daftar kategori produk perusahaan
AF_PELARUT ITEMS	Menyimpan daftar pelarut dari produk perusahaan
AF_VARIAN ITEMS	Menyimpan daftar varian produk perusahaan
AF_WARNA ITEMS	Menyimpan daftar warna dari produk perusahaan
Promosi	Menyimpan data promosi (dalam bentuk excel)

Perusahaan menggunakan aplikasi ERP untuk membantu mengelola kegiatan operasionalnya. Kegiatan operasional tersebut secara garis besar dapat digolongkan menjadi transaksi penjualan produk dan transaksi pembelian *raw material*. Sistem untuk penjualan produk akan mencatat pesanan dari calon pelanggan. Selanjutnya, staf bagian *finance* akan memeriksa riwayat kredit dari calon pelanggannya. Apabila pelanggan tidak memiliki kredit bermasalah, pesanan akan disetujui dan dicatat sebagai *sales order*. Setelah itu, staf gudang akan mengambil dan mengemas barang sesuai pesanan untuk dikirim ke alamat tujuan. Staf gudang akan membuat *packing slip* sesudah barang dikirim. Kemudian, staf *finance* membuat *invoice* setelah mengonfirmasi pengiriman barang. Sementara untuk sistem pembelian *raw material*, ketika *raw material* sudah mendekati nilai minimum stok, maka staf gudang akan membuat dokumen permintaan barang yang ditujukan kepada staf *purchasing*. Staf *purchasing* kemudian membuat dokumen penawaran barang untuk disampaikan ke manajer *purchasing*. *Purchase order* akan dibuat sesudah penawaran barang disetujui oleh manajer *purchasing*.

3.2 Proses ETL

Proses ETL ditujukan untuk membuat tabel fakta dan tabel dimensi. Tabel DATE_DIM dihasilkan dengan melakukan *generate* tanggal dimulai dari 1 Januari 2011, tanggal terlama yang digunakan dalam database. Tabel fakta SALES_FACT, yang digunakan untuk melihat transaksi penjualan produk-produk perusahaan, menggabungkan data master dari tabel SALESTABLE dengan data detail dari tabel SALESLINE. Tabel fakta PACKING_FACT, yang digunakan untuk melihat kegiatan pengiriman barang, menggabungkan data master dari tabel CUSTPACKINGSLIPJOUR dengan data detail dari tabel CUSTPACKINGSLIPTRANS. Tabel fakta INVOICE_FACT, yang digunakan untuk melihat tagihan perusahaan kepada konsumen, menggabungkan data master dari tabel CUSTINVOICEJOUR dengan data detail dari tabel CUSTINVOICETRANS. Tabel fakta PURCHASE_FACT, yang digunakan untuk melihat transaksi pembelian bahan mentah, menggabungkan data master dari tabel PURCHTABLE dengan data detail dari tabel PURCHLINE. Atribut yang sama-sama dimiliki oleh tabel data master dan tabel detail hanya dipakai salah satunya yaitu atribut dari tabel detail. Kemudian, melakukan *extract* ITEM_DIM dan DATE_DIM setiap kali membentuk tabel fakta guna mengganti kode item dan semua atribut bertipe *date* dengan kode dari ITEM_DIM dan DATE_DIM.

Tabel dimensi ITEM_DIM menggabungkan data dari master tabel item, yaitu tabel ECORESPRODUCT, dengan tabel detail item, yaitu tabel INVENTTABLE. Atribut yang sama-sama dimiliki kedua tabel hanya dipakai salah satunya yaitu atribut dari tabel detail item. Selain itu, tabel ini akan di*join* dengan tabel AF_BRANDITEMS, AF_GOLONGANITEMS, AF_JENISITEMS, AF_KATEGORIITEMS, AF_PELARUTITEMS, AF_VARIANITEMS, dan AF_WARNAITEMS untuk melengkapi nama brand, nama golongan, nama jenis, nama kategori, nama pelarut, nama varian, dan nama warna dari setiap item. Tabel ITEM_DIM juga diperlengkapi dengan atribut DATE_START, DATE_END, dan VERSION supaya mudah diketahui riwayat perubahan datanya.

3.3 Proses Analisis Faktor Penentu dengan *LASSO Regression*

Langkah-langkah untuk melakukan analisis faktor menggunakan metode *LASSO regression* adalah sebagai berikut:

1. Data yang dibutuhkan akan diambil dari tabel sales_fact, packing_fact, invoice_fact, item_dim, dan date_dim. Agar proses analisis dapat dilakukan lebih efisien, atribut dari data yang tidak diperlukan untuk pengujian akan di*drop*.

2. Mengolah data tabel INVENTSUM untuk mengambil data jumlah stok barang yang tersedia pada saat konsumen melakukan transaksi beli barang tersebut.
3. Mengolah data untuk mengambil jumlah pendapatan selama satu tahun dari *brand* produk pada tahun transaksi.
4. Mengurangi tanggal SHIPPINGDATEREQUESTED terhadap DELIVERYDATE untuk mendapatkan selisih hari waktu pengiriman barang dari waktu yang diinginkan konsumen.
5. Melakukan standardisasi terhadap faktor-faktor yang akan diuji.
6. Membagi data menjadi *training dataset* dan *testing dataset*.
7. Melakukan *tuning* parameter λ (alpha) dengan *cross validation*.
8. Menggunakan λ (alpha) terpilih untuk membentuk model faktor dari metode *LASSO regression*.
9. Mengukur RMSE, R-squared, dan VIF dari model faktor yang terbentuk.

3.4 Proses Analisis Faktor Penentu dengan *Gradient Boosted Regression Tree*

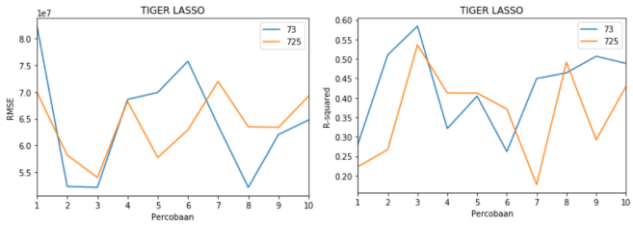
Langkah-langkah untuk melakukan analisis faktor menggunakan metode *Gradient Boosted Regression Tree* adalah sebagai berikut:

1. Data yang dibutuhkan akan diambil dari tabel sales_fact, packing_fact, invoice_fact, item_dim, dan date_dim. Agar proses analisis dapat dilakukan lebih efisien, atribut dari data yang tidak diperlukan untuk pengujian akan di*drop*.
2. Mengolah data tabel INVENTSUM untuk mengambil data jumlah stok barang yang tersedia pada saat konsumen melakukan transaksi beli barang tersebut.
3. Mengolah data untuk mengambil jumlah pendapatan selama satu tahun dari *brand* produk pada tahun transaksi.
4. Mengurangi tanggal SHIPPINGDATEREQUESTED terhadap DELIVERYDATE untuk mendapatkan selisih hari waktu pengiriman barang dari waktu yang diinginkan konsumen.
5. Melakukan standardisasi terhadap faktor-faktor yang akan diuji.
6. Membagi data menjadi *training dataset* dan *testing dataset*.
7. Melakukan *tuning* parameter *learning rate*, *n_estimators*, dan *max_depth* dengan *cross validation*.
8. Menggunakan parameter yang terpilih terpilih untuk membentuk model faktor dari metode *Gradient Boosted Regression Tree*.
9. Mengukur RMSE, R-squared, dan VIF dari model faktor yang terbentuk.

4. PENGUJIAN SISTEM

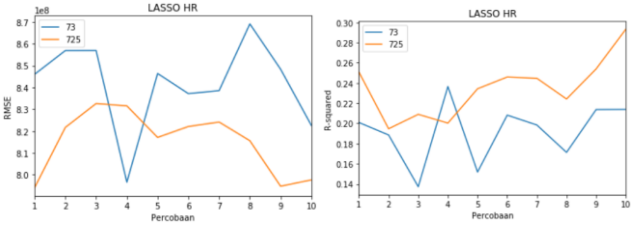
4.1 Pemilihan Rasio *Training Dataset* dan *Testing Dataset*

Dataset dibagi menjadi *training dataset* dan *testing dataset* untuk menguji kemampuan model faktor dalam memprediksi data diluar data yang dipakai untuk membentuk model faktor tersebut. Pada penelitian *LASSO regression* [8] digunakan 70% sebagai *training dataset* dan 30% sebagai *testing dataset* sedangkan pada penelitian *Gradient Boosted Regression Tree* [10] menggunakan 75% sebagai *training dataset* dan 25% sebagai *testing dataset*. Karena itu, pengujian dilakukan dengan membentuk model faktor dari 3 *brand* menggunakan *LASSO regression* dan *Gradient Boosted Regression Tree* dengan kedua rasio. Kode '73' menandakan penggunaan rasio 70:30 sementara kode '725' menandakan penggunaan rasio 75:25.



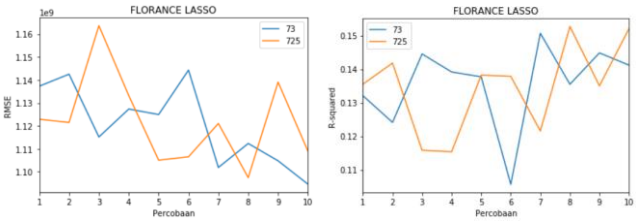
Gambar 2. Model faktor *brand* TIGER dengan LASSO

Pada Gambar 2, dari 10 kali percobaan membentuk model faktor *brand* TIGER dengan *LASSO regression*, penggunaan rasio 70:30 mampu menghasilkan model faktor terbaik dengan nilai R-squared tertinggi yaitu 0.5841992684112578 dan RMSE paling rendah yaitu 52168970.12143629 pada percobaan ke 3.



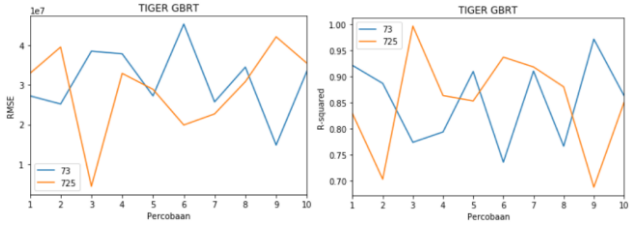
Gambar 3. Model faktor *brand* HR dengan LASSO

Pada Gambar 3, dari 10 kali percobaan membentuk model faktor *brand* HR dengan *LASSO regression*, penggunaan rasio 70:30 pada percobaan ke 4 memiliki tingkat *error* (RMSE) yang cukup rendah tetapi kemampuan prediksinya juga rendah dilihat dari nilai R-squared. Oleh karena itu, penggunaan rasio 75:25 lebih baik ketika membentuk model faktor *brand* HR.



Gambar 4. Model faktor *brand* FLORANCE dengan LASSO

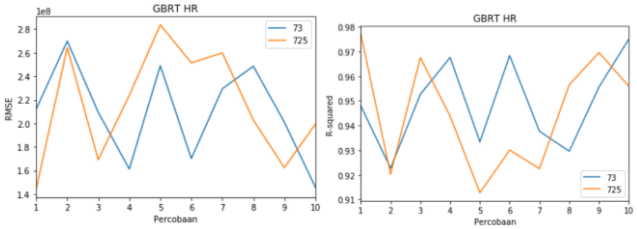
Pada Gambar 4, dari 10 kali percobaan membentuk model faktor *brand* FLORANCE dengan *LASSO regression*, tingkat *error* terendah dilihat dari nilai RMSE dihasilkan pada percobaan ke 10 dengan perbandingan 70:30 namun memiliki nilai R-squared yang lebih rendah dari model faktor dengan perbandingan 75:25. Oleh sebab itu, model faktor pada percobaan ke 8 menggunakan perbandingan 75:25 lebih baik dengan nilai RMSE 1097447141.0271187 dan R-squared sebesar 0.15269617543141611.



Gambar 5. Model faktor *brand* TIGER dengan GBRT

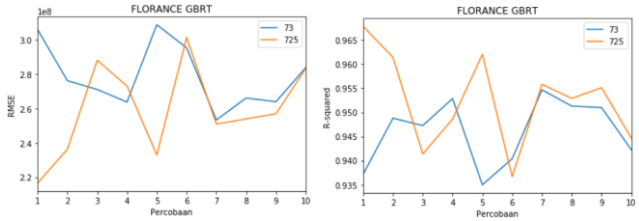
Gambar 5 menunjukkan dari 10 kali percobaan membentuk model faktor *brand* TIGER dengan metode GBRT, penggunaan rasio 75:25 mampu menghasilkan model faktor terbaik dengan nilai R-

squared tertinggi yaitu 0.9969398115580738 dan RMSE paling rendah yaitu 4387392.707828254 pada percobaan ke 3.



Gambar 6. Model faktor *brand* HR dengan GBRT

Gambar 6 menunjukkan dari 10 kali percobaan membentuk model faktor *brand* HR dengan metode GBRT, penggunaan rasio 75:25 mampu menghasilkan model faktor terbaik pada percobaan ke 1 dengan nilai R-squared dan RMSE sedikit lebih baik daripada model faktor dengan rasio 70:30 pada percobaan ke 10. Model faktor yang pertama dengan R-squared yaitu 0.9770802276730948 dan nilai RMSE yaitu 144267249.5189476 sedangkan percobaan ke 10 memiliki nilai RMSE sebesar 145783476.8266843 dan 0.9750008048507338 untuk nilai R-squared.

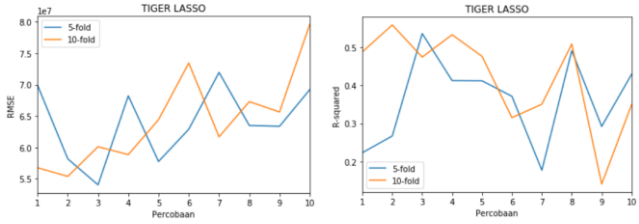


Gambar 7. Model faktor *brand* FLORANCE dengan GBRT

Gambar 7 menunjukkan dari 10 kali percobaan membentuk model faktor *brand* FLORANCE dengan metode GBRT, model faktor dengan perbandingan 75:25 menunjukkan hasil terbaik pada percobaan ke 1 dengan nilai R-squared sebesar 0.9678328161646199 dan RMSE sebesar 216599369.70230198. Berdasarkan semua percobaan ketiga brand dipilih perbandingan *training dataset* dan *testing dataset* yaitu 75:25 untuk penelitian ini.

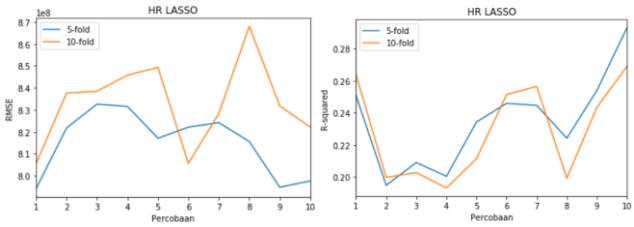
4.2 Pemilihan *K-fold Cross Validation*

Parameter terbaik dari metode *LASSO regression* dan *Gradient Boosted Regression Tree* ditentukan dengan *cross validation*. Penelitian *LASSO regression* [8] menggunakan *10-fold* sedangkan penelitian *Gradient Boosted Regression Tree* [10] menggunakan *5-fold*. Oleh karena itu, pengujian dilakukan dengan membentuk model faktor dari 3 *brand* menggunakan *LASSO regression* dan *Gradient Boosted Regression Tree* untuk memilih *k-fold* yang bisa menghasilkan model faktor terbaik.



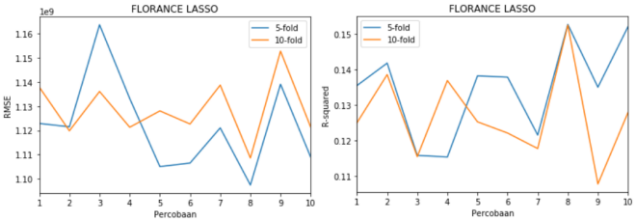
Gambar 8. *Cross validation brand* TIGER dengan LASSO

Pada Gambar 8, dari 10 kali percobaan membentuk model faktor *brand* TIGER dengan *LASSO regression*, RMSE terbaik didapat pada percobaan ke 3 menggunakan 5-*fold* dengan nilai 54027180.391916506 dan nilai R-squared 0.5359553627507709. Namun, R-squared tersebut masih lebih rendah dibandingkan model faktor pada percobaan kedua menggunakan 10-*fold* yang memiliki nilai R-squared tertinggi yaitu 0.5585115633314329.



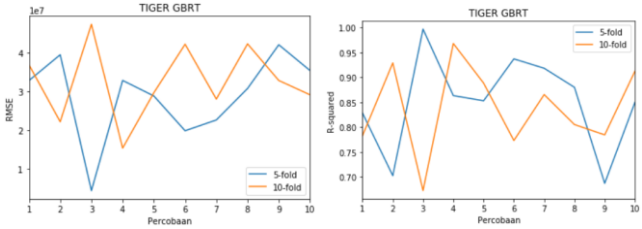
Gambar 9. *Cross validation brand* HR dengan LASSO

Pada Gambar 9, dari 10 kali percobaan membentuk model faktor *brand* HR dengan *LASSO regression*, didapat model faktor pada percobaan ke 9 dan percobaan ke 10 yang menggunakan 5-*fold cross validation* menghasilkan RMSE dan R-squared terbaik.



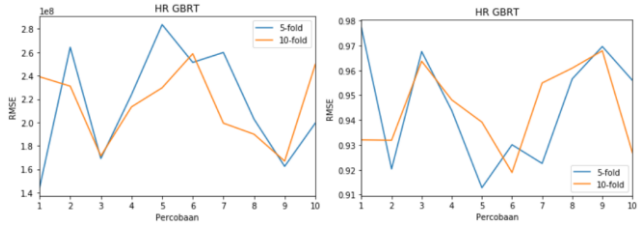
Gambar 10. *Cross validation* FLORANCE dengan LASSO

Pada Gambar 10, dari 10 kali percobaan membentuk model faktor *brand* FLORANCE dengan *LASSO regression*, model faktor menggunakan 5-*fold cross validation* pada percobaan ke 8 menunjukkan R-squared yang terbaik diikuti model faktor pada percobaan ke 10. Percobaan ke 8 menggunakan 5-*fold* memiliki nilai R-squared sebesar 0.15269617543141611 sedikit lebih tinggi daripada model faktor yang menggunakan 10-*fold* dengan nilai R-squared 0.15248583493555523.



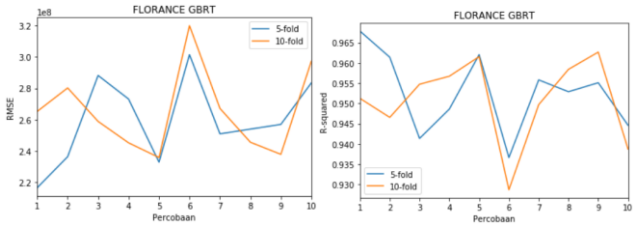
Gambar 11. *Cross validation brand* TIGER dengan GBRT

Gambar 11 menunjukkan dari 10 kali percobaan membentuk model faktor *brand* TIGER dengan metode GBRT, 5-*fold cross validation* menghasilkan model faktor terbaik dengan nilai R-squared tertinggi yaitu 0.9969398115580738 dan RMSE paling rendah yaitu 4387392.707828254 pada percobaan ke 3.



Gambar 12. *Cross validation brand* HR dengan GBRT

Sedangkan, Gambar 12 menunjukkan dari 10 kali percobaan membentuk model faktor *brand* HR dengan metode GBRT, model faktor yang menggunakan 5-*fold cross validation* mengungguli model faktor yang menggunakan 10-*fold cross validation* pada percobaan ke 1, percobaan ke 3, dan percobaan ke 9.



Gambar 13. *Cross validation* FLORANCE dengan GBRT

Gambar 13 menunjukkan dari 10 kali percobaan membentuk model faktor *brand* FLORANCE dengan metode GBRT, 5-*fold cross validation* menghasilkan model faktor terbaik dengan nilai R-squared sebesar 0.9678328161646199 dan RMSE sebesar 216599369.70230198 pada percobaan ke 1. Berdasarkan semua percobaan ketiga *brand* dipilih 5-*fold cross validation* untuk digunakan pada penelitian ini.

4.3 Pengukuran Model Faktor

Model faktor dibentuk untuk brand-brand perusahaan PT. X yang tingkat penjualannya tidak memuaskan selama beberapa tahun belakangan. Kemudian, data dibagi menjadi 75% sebagai *training dataset* dan 25% sebagai *testing dataset*. Salah satu faktor, yaitu durasi promosi, tidak dapat diuji karena data promosi tidak dapat diasosiasikan dengan data transaksi penjualan. Parameter *alpha* yang tepat dari metode *LASSO regression* juga parameter *learning_rate*, *n_estimators*, dan *max_depth* dari *Gradient Boosted Regression Tree* ditentukan menggunakan *cross validation* dengan pengukuran *mean squared error*.

Lalu, model faktor dari *LASSO regression* dan *Gradient Boosted Regression Tree* diukur dengan *Root Mean Squared Error*, *R-squared*, dan *Variance Inflation Factor*. Pengukuran model faktor bertujuan untuk mengetahui faktor-faktor yang berpengaruh signifikan terhadap penjualan dari masing-masing *brand*.

4.3.1 Brand TIGER

Model faktor yang dibentuk dengan metode GBRT menunjukkan nilai RMSE yang lebih rendah, yaitu 32845494.843187317, diikuti R-squared yang lebih tinggi, yaitu 0.8284910568298998, dibandingkan model faktor dari metode *LASSO regression* yang memiliki nilai RMSE sebesar 69905470.93658175 dan nilai R-squared sebesar 0.22311366242937725. Nilai VIF dari variabel-variabel yang membentuk model faktor berada dibawah 4 menunjukkan tingkat *multicollinearity* yang rendah [3].

4.3.2 Brand QQ

Model faktor yang dibentuk menggunakan metode GBRT menunjukkan nilai RMSE yang lebih rendah, sebesar 199041580.61365184, diikuti R-squared yang lebih tinggi, sebesar 0.8941438605100821, jika dibandingkan dengan model faktor dari metode *LASSO regression* yang memiliki nilai RMSE 561105303.2366432 dan nilai R-squared hanya 0.15876530745760578. Nilai VIF dari variabel-variabel yang membentuk model faktor berada dibawah 4 menunjukkan rendahnya *multicollinearity*.

4.3.3 Brand HR

Model faktor yang dibentuk dengan metode GBRT menunjukkan nilai RMSE lebih rendah, yaitu 144267249.5189476, diikuti dengan nilai R-squared lebih tinggi, yaitu 0.9770802276730948, jika dibandingkan dengan model faktor dari metode *LASSO regression* yang memiliki nilai RMSE sebesar 794151683.1375011 dan nilai R-squared sebesar 0.25079977445244417. Nilai VIF dari variabel-variabel yang membentuk model faktor berada dibawah 4 menunjukkan rendahnya *multicollinearity*.

4.3.4 Brand PUMA

Model faktor yang dibentuk dengan metode GBRT menunjukkan nilai RMSE yang lebih rendah, yaitu 549729488.3220727, diikuti dengan R-squared yang lebih tinggi, yaitu 0.9536064874860142, jika dibandingkan dengan model faktor dari metode *LASSO regression* dengan nilai RMSE sebesar 2342715340.7803597 dan R-squared sebesar 0.15744505846363843. Nilai VIF dari variabel-variabel yang membentuk model faktor berada dibawah 4 menunjukkan rendahnya *multicollinearity*.

4.3.5 Brand MASTER PLAMIR

Model faktor yang dibentuk dengan metode GBRT menunjukkan nilai RMSE sebesar 207265903.0237471 dan nilai R-squared sebesar 0.934456271060286. Model faktor tersebut memiliki nilai RMSE lebih rendah dan nilai R-squared lebih tinggi daripada model faktor dari metode *LASSO regression* dengan nilai RMSE sebesar 802675129.3426766 dan R-squared sebesar 0.08983364587111276. Nilai VIF dari variabel-variabel yang membentuk model faktor berada dibawah 4 menunjukkan rendahnya *multicollinearity*.

4.3.6 Brand TAKA ARMOR

Model faktor yang dibentuk dengan metode GBRT menunjukkan nilai RMSE sebesar 513026012.9063656 dan nilai R-squared sebesar 0.8582592696870198. Model faktor tersebut memiliki nilai RMSE lebih rendah dan nilai R-squared lebih tinggi daripada model faktor dari metode *LASSO regression* dengan nilai RMSE sebesar 1297352968.8097699 dan R-squared sebesar 0.12211264610266061. Nilai VIF dari variabel-variabel yang membentuk model faktor brand ini berada dibawah 4 yang menunjukkan rendahnya *multicollinearity*.

4.3.7 Brand FLORANCE

Model faktor yang dibentuk dengan metode GBRT menunjukkan nilai RMSE yang lebih rendah, yaitu 216599369.70230198, diikuti dengan nilai R-squared yang lebih tinggi, yaitu 0.9678328161646199, jika dibandingkan model faktor dari metode *LASSO regression* yang memiliki nilai RMSE sebesar 1122876813.1080298 dan R-squared hanya 0.13550419151343096. Nilai VIF dari variabel-variabel yang membentuk model faktor brand ini berada dibawah 4 yang menunjukkan rendahnya *multicollinearity*.

4.3.8 Brand TAKA IMPREZZA

Model faktor yang dibentuk dengan metode GBRT menunjukkan nilai RMSE yang lebih rendah, yaitu 544116016.2447563, diikuti dengan R-squared yang lebih tinggi, yaitu 0.884915069002122, daripada model faktor dari metode *LASSO regression* dengan nilai RMSE sebesar 1460248093.2995691 dan nilai R-squared sebesar 0.15333031497023908. Nilai VIF dari variabel-variabel yang membentuk model faktor brand ini berada dibawah 4 yang menunjukkan rendahnya *multicollinearity*.

4.3.9 Brand SUPER

Model faktor yang dibentuk dengan metode GBRT menunjukkan nilai RMSE yang lebih rendah, yaitu 3549669531.61878, diikuti dengan R-squared yang lebih tinggi, yaitu 0.9585511460610343, daripada model faktor dari metode LASSO regression dengan nilai RMSE sebesar 16822713663.759274 dan nilai R-squared sebesar 0.06904478944165182. Nilai VIF dari variabel costprice dan salesprice yang membentuk model faktor brand yaitu 4 lebih sedikit masih tergolong tingkat multicollinearity yang rendah [3].

4.3.10 Brand MAWAR

Model faktor yang dibentuk menggunakan metode GBRT menunjukkan nilai RMSE yang lebih rendah, yaitu 176269781.04350525, diikuti dengan R-squared yang lebih tinggi, yaitu 0.981450627001505, daripada model faktor dari metode *LASSO regression* dengan nilai RMSE sebesar 1256542282.7932003 dan nilai R-squared hanya 0.03840063021532358. Nilai VIF dari variabel-variabel yang membentuk model faktor brand ini berada dibawah 4 yang menunjukkan rendahnya *multicollinearity*.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil analisis pada pengujian sistem, dapat diambil beberapa kesimpulan antara lain:

- Analisis faktor penentu yang dihasilkan melalui metode *Gradient Boosted Regression Tree* berhasil memilih faktor-faktor yang berpengaruh signifikan terhadap penjualan PT. X dari antara faktor-faktor yang diuji dengan tingkat akurasi yang lebih tinggi daripada metode *LASSO regression*.
- Faktor-faktor penentu, yang merupakan hasil analisis dengan metode Gradient Boosted Regression Tree, untuk penjualan produk brand TIGER, TAKA ARMOR, FLORANCE, TAKA IMPREZZA, SUPER, dan MAWAR adalah costprice, salesprice, quantity, stock, bulan pemesanan, dan ketepatan pengiriman. Sedangkan, faktor-faktor penentu untuk penjualan produk brand QQ, HR, PUMA, dan MASTER PLAMIR adalah costprice, salesprice, quantity, bulan pemesanan, dan ketepatan pengiriman.

5.2 Saran

Saran untuk melakukan pengembangan lebih lanjut adalah sebagai berikut:

- Analisis dilakukan untuk menguji faktor-faktor lain yang memiliki pengaruh terhadap penjualan produk-produk PT. X baik secara langsung maupun tidak langsung seperti persentase diskon yang diberikan, tipe pengiriman barang, jenis transaksi.
- Melakukan analisis untuk mengetahui relevansi dari analisis faktor penentu yang dihasilkan dalam mempengaruhi penjualan PT. X di masa yang akan datang.

6. DAFTAR REFERENSI

[1] Bel, G. and Warner, M. E. 2015. Factors explaining inter-municipal cooperation in service delivery: a meta-regression analysis. *Journal of Economic Policy Reform* 19, 2, 91–115. DOI= https://doi.org/10.1080/17487870.2015.1100084.

[2] Ferreira, J. C., De Almeida, J., and Da Silva, A. R. 2015. The Impact of Driving Styles on Fuel Consumption: A Data-Warehouse-and-Data-Mining-Based Discovery Process. *IEEE Transactions on Intelligent Transportation Systems* 16, 5, 2653–2662. DOI= https://doi.org/10.1109/TITS.2015.2414663.

[3] Garcia, R., Diaz, G., Pañeda, X. G., Tuero, A. G., Pozueco, L., Melendi, D., Sanchez, J. A., Corcoba, V., and Pañeda, A. G. 2017. Impact of efficient driving in professional bus fleets. *Energies* 10, 12, 1–25. DOI= https://doi.org/10.3390/en10122060.

[4] Gawande, S. 2015. 3 Reasons Why You Need to Perform ETL Testing. URI= https://icedq.com/etl-testing/3-reasons-why-you-need-to-perform-etl-testing.

[5] Hepp, T., Schmid, M., Gefeller, O., Waldmann, E., and Mayr, A. 2016. Approaches to regularized regression - A comparison between gradient boosting and the lasso. *Methods of Information in Medicine* 55, 5, 422–430. DOI= https://doi.org/10.3414/ME16-01-0033.

[6] Linstedt, D., and Olschimke, M. 2015. *Building a Scalable Data Warehouse with Data Vault 2.0*. Elsevier.

[7] McNeish, D. M. 2015. Using Lasso for Predictor Selection and to Assuage Overfitting: A Method Long Overlooked in Behavioral Sciences. *Multivariate Behavioral Research* 50, 5, 471–484. DOI= https://doi.org/10.1080/00273171.2015.1036965.

[8] Mueller-Using, S., Feldt, T., Sarfo, F. S., and Eberhardt, K. A. 2016. Factors associated with performing tuberculosis screening of HIV-positive patients in Ghana: LASSO-based predictor selection in a large public health data set. *BMC Public Health* 16, 1, 1–8. DOI= https://doi.org/10.1186/s12889-016-3239-y.

[9] Naeem, T. 2020. Data Warehouse Concepts: Kimball vs. Inmon Approach. URI= https://www.astera.com/type/blog/data-warehouse-concepts/.

[10] Persson, C., Bacher, P., Shiga, T., and Madsen, H. 2017. Multi-site solar power forecasting using gradient boosted regression trees. *Solar Energy* 150, 423– 436. DOI= https://doi.org/10.1016/j.solener.2017.04.066.

[11] Shin, Y. 2015. Application of boosting regression trees to preliminary cost estimation in building construction projects. *Computational Intelligence and Neuroscience* 2015, 9 pages. DOI= https://doi.org/10.1155/2015/149702.

[12] Smallcombe, M. 2019. The Ultimate Guide to Data Warehouse Design. URI= https://www.xplenty.com/blog/the-ultimate-guide-to-data-warehouse-design/.

[13] Sreemathy, J., Priyadharshini, S., Radha, K., Sangeerna, K., and Nivetha, G. 2019. Data Validation in ETL Using TALEND. *2019 5th International Conference on Advanced Computing and Communication Systems, ICACCS 2019*, 1183– 1186. DOI= https://doi.org/10.1109/ICACCS.2019.8728420.

[14] Yu, W., Zhao, C., Wu, H., and Peng, C. 2019. Analysis of Vegetable Price Fluctuation Law and Causes based on Lasso Regression Model. *Journal of Physics: Conference Series* 1284, 1. DOI= https://doi.org/10.1088/1742-6596/1284/1/012002.