Komparasi Model Klasifikasi untuk Segmentasi Pelanggan pada Perusahaan Retail TagMedia

Gladys Patricia
Information Systems
Universitas Multimedia Nusantara
Banten, Indonesia
gladys.patricia@student.umn.ac.id

***Abstract*—TagMedia is a retail company for office and school supply products. In order to improve business performance, TagMedia needs to understand their customer better and treat them specifically. This research is conducted to predict customer value segment based on profit, recency, sales, and shopping frequency using several algorithms. The methods used is based on CRISP-DM framework. All of the models will be compared based on validation misclassification rate. The results show that Decision Tree has a value of 0.0528, Neural Network of 0.0906, Random Forest of 0.0415, and Gradient Boosting of 0.0302. The misclassification value is expected to be as minimal as possible to get an optimal model. Therefore, Gradient Boosting is selected. TagMedia can take advantage of this model to predict customer value more accurately and treat their customer better.**

Keywords—Analysis, Classification, Decision Tree, Neural Network, Gradient Boost, Random Forest, Prediction

I. LATAR BELAKANG & PEMAHAMAN BISNIS

Retail merupakan aktivitas bisnis untuk menjual barang atau jasa kepada pengguna akhir. Umumnya, pembeli produk retail tidak menjual kembali barang tersebut, melainkan digunakan untuk keperluan pribadi, keluarga, atau kebutuhan sehari-hari lainnya [1]. Seiring dengan perkembangan zaman, perusahaan retail semakin marak untuk ditemui. Hal ini didukung oleh jumlah *retailer* di seluruh dunia mencapai angka 9.1 juta [2]. Secara singkat, bisnis dijalankan untuk mendapatkan keuntungan. Dalam kondisi persaingan pasar yang ketat, sebuah bisnis perlu menyusun strategi dengan baik sehingga unggul dibandingkan dengan kompetitor. Pada saat inilah bisnis perlu didukung oleh strategi analisis data sehingga menghasilkan informasi yang berharga sebagai acuan untuk pergerakan bisnis.

Saat ini, penggunaan ilmu data dalam bisnis sedang berada dalam tren. Perusahaan di seluruh dunia mencoba untuk mendapatkan manfaat dari data-data pelanggan yang mereka miliki, salah satunya untuk meningkatkan pendapatan bisnis [3]. Melalui pengolahan data, perusahaan terbantu dalam memahami pasar dengan lebih baik dan membuat keputusan dengan tepat dan lebih akurat [4]. Perusahaan atau bisnis dalam sektor apapun dapat memanfaatkan analisis data. Data yang digunakan dapat berasal dari sumber *internal* maupun *eksternal* seperti sosial media. Pengolahan ini membantu perusahaan untuk mendapatkan *competitive advantage* dari *insight* yang diperoleh [5].

Penelitian ini berfokus pada kasus di perusahaan retail yang menjual barang-barang perlengkapan alat kantor dan sekolah (*office supply*). Perusahaan ini bernama TagMedia atau dikenal juga dengan nama TagMe. TagMedia bertindak sebagai *retailer* yang menjual kembali produk dari berbagai *supplier*. Pendapatan TagMedia diperoleh dari keuntungan selisih harga jual dengan harga beli setiap produk. TagMedia memiliki toko fisik yang berpusat di Amerika Serikat dan melakukan penjualan *online* melalui *website* resmi TagMedia serta *platform e-commerce* seperti Amazon. Selama menjalankan bisnisnya, TagMedia belum menggunakan data untuk mengambil keputusan bisnis. TagMedia beberapa kali mengalami kerugian dalam transaksi penjualan produknya. Selain itu juga mengalami kesulitan untuk menargetkan produk kepada pelanggan. Dengan demikian, perlu dilakukan analisis data terhadap kondisi TagMedia saat ini dan mengambil informasi-informasi untuk membantu pengembangan bisnis.

II. TINJAUAN TEORITIS

A. Penelitian Terdahulu

Pada kasus TagMedia, akan dilakukan klasifikasi terhadap segmentasi pelanggan menggunakan beberapa algoritma. Klasifikasi dilakukan untuk memahami pelanggan TagMedia sehingga dapat menawarkan produk yang relevan dengan minat dari pelanggan. Terdapat beberapa penelitian terdahulu mengenai topik yang serupa sebagai bahan acuan dan pembelajaran dari penelitian ini, yaitu:

- Penelitian oleh Pratomo, Najib, dan Mulyati tahun 2019 dengan judul "Customer Segmentation Analysis based on The Customer Lifetime Value Method" yang membagi pelanggan menjadi tiga kategori besar berdasarkan *value* yang diberikan untuk perusahaan [6]. Penelitian ini membagi *value* pelanggan (CLV) ke dalam 4 kategori berbeda, yaitu *high, mid, low, very low*.
- Penelitian oleh Choudhury dan Nur tahun 2019 dengan judul "A Machine Learning Approach to Identify Potential Customer Based on Purchase Behavior" [7]. Penelitian ini membagi pelanggan menjadi dua kelas, yaitu *high potential* and *low potential customer*. Prediksi dilakukan dengan 5 algoritma yang berbeda untuk mendapatkan performa terbaik, yaitu *Logistic Regression, Decision Tree Classifier, Suppoer Vector Classifier, Random Forest Classifier*, dan *Multilayer Perceptron Classifier*.
- Penelitian oleh Favian dan Suryani tahun 2020 dengan judul "A Case Study of Applying Customer Segmentation in A Medical Equipment Industry" [8]. Penelitian ini membagi pelanggan perlengkapan medis menjadi beberapa kluster. Kluster tersebut diperoleh melalui variabel *length, recency, frequency,* dan

monetary (LRFM). Kemudian membentuk model prediksi dengan klasifikasi IF-THEN *rules*.

B. CRISP-DM

CRISP-DM (*Cross Industry Standard Process for Data Mining*) merupakan model pengolahan data dalam industri yang digunakan untuk *data mining*. Model ini terdiri atas 6 fase, dimulai dari pemahaman terhadap bisnis hingga implementasi dari hasil pengolahan data tersebut [9]. Pemahaman bisnis dilakukan untuk mengetahui kondisi bisnis saat ini dan apa saja yang perlu diperbaiki. Selanjutnya masuk ke tahap kedua, yaitu pemahaman terhadap data untuk mengetahui apa yang dapat dilakukan dengan data tersebut. Pada fase ketiga, data perlu dipersiapkan dan dibereskan terlebih dahulu agar mendapatkan model yang optimal. Pada tahap pembuatan model, teknik yang dapat digunakan beragam sesuai dengan kebutuhan dan jenis data *input.* Model tersebut kemudian dievaluasi untuk mengetahui performanya dan kelayakannya untuk diimplementasi. Apabila menggunakan lebih dari satu model, maka perlu dilakukan pemilihan model yang paling optimal. Selanjutnya, tahap terakhir adalah penggunaan model dan pengetahuan untuk diimplementasikan sebagai komponen dari *software*.

C. *Exploratory Data Analysis*

Exploratory Data Analysis merupakan tahapan untuk menganalisa, menginvestigasi, dan merangkum karakteristik utama dari suatu dataset. Proses *Exploratory Data Analysis* (EDA) membantu pemahaman terhadap data, mempermudah manipulasi dan pencarian *pattern*, anomali, atau informasi dari data yang dimiliki [10]. Pada penelitian ini, eksplorasi data dilakukan melalui statistika deskriptif dan pembuatan grafik.

Statistika deskriptif merupakan tahapan yang sangat umum dilakukan dalam proses *data science*. Statistika deskriptif dilakukan untuk mengetahui karakteristik dari data dengan mengubahnya menjadi metrik numerik sederhana. Proses eksplorasi data tidak hanya dilakukan melalui statistika deskriptif, tetapi juga dapat dilakukan melalui visualisasi. Visualisasi merupakan teknik untuk merepresentasikan data dalam bentuk diagram, animasi, gambar, dan bentuk visual lainnya untuk menyediakan informasi sehingga lebih mudah dipahami.

D. *Decision Tree*

Decision Tree atau pohon keputusan merupakan salah satu algoritma yang dapat digunakan untuk mengelompokan dan melakukan klasifikasi data. *Decision Tree* berbentuk seperti pohon dan mendeskripsikan setiap kelas agar dapat menemukan pola tertentu dari data. Secara singkat, *Decision Tree* bekerja dengan menghitung nilai *Information Gain* (IG) dari setiap kolom yang ada, memilih *Information Gain* (IG) yang paling besar, dan membentuk *rule* untuk setiap *node* dari atribut yang telah dianalisis. Setiap data memasuki *Decision Tree* mulai dari *node* paling atas hingga daun terakhir. Pada daun terakhir, data tersebut mendapatkan kelas tertentu berdasarkan *rule* yang telah ditetapkan oleh model [11].

E. *Neural Network*

Neural Network merupakan algoritma klasifikasi yang terinspirasi dari cara kerja otak manusia. *Neural Network* terusun dari model yang berbentuk seperti jaringan saraf manusia dan saling terkoneksi untuk membuat keputusan. Berdasarkan konsep tersebut, *Neural Network* memiliki lapisan-lapisan *layer* dengan tingkat kompleksitas bergantung dengan jumlah *input* dan *output* yang diharapkan. Secara umum, proses *Neural Network* tersusun atas tiga tahapan yaitu insisialisasi, aktivasi, *weight training*, dan iterasi untuk klasifikasi. Setiap tahapan ini dilakukan hingga masalah berhasil dipecahkan [12].

F. *Gradient Boost*

Gradient Boosting merupakan model yang digunakan untuk menyelesaikan permasalahan regresi dan klasifikasi. Algortima ini bekerja dengan cara membentuk prediksi yang simpel secara berurutan. Suatu model akan melakukan prediksi dan melengjapi *error* yang terjadi pada model sebelumnya. *Gradient boosting* ini menggabungkan *weak learners* sehingga menciptakan model *strong learner* [13].

G. *Random Forest Classifier*

Random Forest Classifier merupakan model yang menggabungkan berbagai pohon keputusan seperti pada *Decision Tree*. Setiap pohon dalam *Random Forest* menghasilkan prediksi terhadap kelas. Setiap pohon yang ada akan dievaluasi, pemilihan model terbaik didasarkan pada sistem *vote*. Pohon dengan jumlah *vote* terbesar akan dipilih sebagai model untuk prediksi [14].

III. METODOLOGI PENELITIAN

Tujuan utama dari metode penelitian ini adalah memberikan perkiraan prediksi segmentasi *value* pelanggan TagMedia. Proses penelitian memiliki beberapa tahapan seperti yang digambarkan pada *flow* berikut. Pengolahan data secara umum mengikuti tata kerja CRISP-DM (*Cross Industry Standard Process for Data Mining*).

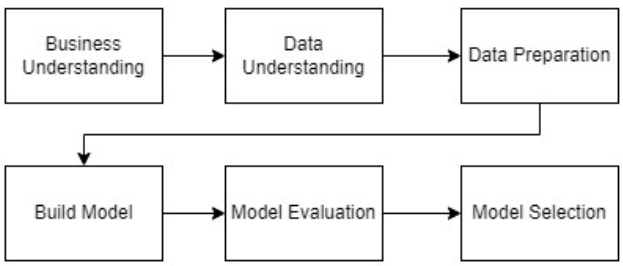


Fig 1. Metodologi Penelitian

A. *Business Understanding*

Tahapan pertama dalam proses *data science* adalah pengenalan terhadap situasi, permasalahan, serta tujuan bisnis. TagMedia merupakan perusahaan retail yang menjual produk peralatan dan perlengkapan kantor. Produk yang dijual terbagi menjadi beberapa kategori seperti kertas, binder, amplop, barang elektronik, dan furnitur perkantoran. Proses bisnis yang dilaksanakan dalam TagMedia dimulai dari pembelian produk dari berbagai *supplier* dengan merk dagang dan jenis produk yang berbeda-beda. Produk tersebut umumnya didapatkan dengan harga yang lebih murah karena dibeli langsung dari tangan pertama. TagMedia kemudian menaikkan harga dari produk dan menjualnya kepada

customer, selisih kenaikan harga tersebut menjadi keuntungan bagi TagMedia. Pelanggan dari TagMedia tidak hanya *customer* perorangan, tetapi juga perusahaan atau perkantoran lain untuk digunakan sendiri (tidak dijual kembali).

Selama menjalankan bisnisnya, TagMedia belum menggunakan data untuk mengambil keputusan bisnis. TagMedia mengalami kesulitan untuk menargetkan produk kepada pelanggan dan tidak memahami pelanggannya sendiri. TagMedia beberapa kali mengalami kerugian atas penjualan produknya. Dengan demikian, perlu dilakukan analisis data dan prediksi terhadap pelanggan berdasarkan tingkatan *value* untuk menargetkan promosi serta memberikan pelayanan yang bersifat *personalized*. Layanan khusus tersebut dilakukan untuk menjaga loyalitas pelanggan dan menarik minat belanja pelanggan TagMedia.

B. *Data Understanding*

Data yang digunakan dalam analisis ini adalah data pelanggan TagMedia yang terdiri atas 19 kolom dan 886 baris. Data ini terdiri atas variabel seperti id pelanggan, nama, usia, lokasi, *recency*, frekuensi atau banyaknya transaksi, jumlah nilai *sales* dalam dolar yang dilakukan terhadap pelanggan tersebut, jumlah *profit*, dan *value segment*. *Recency* merupakan jarak pembelian seorang pelanggan dalam satuan hari, sedangkan *value segment* merupakan variabel berupa "label" atas pelanggan. *Value segment* ini menentukan apakah pelanggan tersebut memberikan nilai yang rendah, sedang, atau tinggi bagi perusahaan.

Berdasarkan data tersebut akan dilakukan prediksi terhadap tipe *value* pelanggan bagi TagMedia. *Value segment* berperan sebagai target, sedangkan variabel lainnya merupakan prediktor yang akan menyusun *rule* untuk klasifikasi. Data sejumlah 886 baris ini akan dipecah menjadi data *training* dan *testing* untuk pembentukan dan validasi model.

C. *Data Preparation*

Data preparation merupakan tahap persiapan data sebelum memasuki tahap pembuatan model. Dalam persiapan data, umumnya dilakukan pembersihan terhadap format, *missing value*, penambahan fitur, pemilihan fitur, dan lain-lain. Data pelanggan TagMedia telah ada dalam bentuk yang terstruktur dan memiliki cukup kolom untuk analisis, sehingga tidak memerlukan banyak persiapan data.

Hal lain yang penting dalam proses ini adalah *Exploratory Data Analysis* (EDA) yang digunakan untuk memahami data dan menemukan pola-pola atau anomali yang tersembunyi. Pada kasus TagMedia, EDA dilakukan melalui visualisasi data pelanggan.

D. *Modeling*

Tahap selanjutnya setelah persiapan dan eksplorasi data adalah pembuatan model. Pada tahap ini, model prediksi klasifikasi pelanggan dilakukan menggunakan beberapa algoritma, yaitu *Decision Tree*, *Neural Network*, *Gradient Boosting*, dan *Random Forest Classifier*. Keempat algoritma tersebut akan mendapatkan *input* yang sama dan diharapkan untuk menghasilkan klasifikasi sebaik mungkin. Model dilatih menggunakan data *training* dan divalidasi melalui data *testing* dengan partisi 70:30.

E. *Model Evaluation*

Evaluasi model merupakan tahap pengukuran performa dari algoritma yang digunakan. Meskipun *input* yang diberikan sama, namun setiap algoritma bekerja dengan cara yang berbeda-beda, sehingga hasil prediksinya dapat bervariasi. Evaluasi ini dilakukan dengan melihat angka *validation misclassification rate* pada SAS Visual Analytics. Semakin kecil nilai kesalahan klasifikasi, maka model tersebut memiliki performa yang semakin baik.

F. *Model Selection*

Seleksi model dilakukan untuk memilih model terbaik dari berbagai algoritma yang digunakan. Model dengan nilai *misclassification rate* terkecil akan dipilih untuk masuk ke tahap *deployment* untuk kebutuhan prediksi selanjutnya oleh TagMedia.

IV. HASIL DAN DISKUSI

Proses pengolahan data dilakukan menggunakan SAS Data Studio, SAS Studio, dan SAS Visual Analytics dengan *input* berupa data pelanggan TagMedia. Seluruh fasilitas tersebut digunakan untuk membersihkan data, mendeskripsikan data, dan membuat model. Berikut ini merupakan hasil dari setiap tahap penelitian:

A. *Data Preparation*

Persiapan data dilakukan melalui SAS Data Studio (*prepare data*) untuk menghapus data-data yang tidak digunakan untuk analisis. Data yang dihapus adalah *frequency score, recency score, sales score, profit score,* dan *overall score* yang sebenarnya hanya berisi data label dalam rentang 1-4 (lemah-tinggi). Label ini memiliki nilai dan makna yang sama dengan variabel asli (*frequency*, *sales*, *profit*, dan *recency*), dengan demikian dilakukan penghapusan. Hasil akhirnya adalah sebagai berikut, tersisa 14 variabel untuk dianalisis dan pembuatan model:

#	Name	Label	Type	Raw Length	Formatted L...
1	CustomerID		varchar	8	8
2	CustomerName		varchar	22	22
3	Gender		varchar	6	6
4	Age		double	8	12
5	CustomerType		varchar	11	11
6	City		varchar	17	17
7	State		varchar	14	14
8	PostalCode		double	8	12
9	Region		varchar	7	7
10	Recency		double	8	12
11	Frequency		double	8	12
12	Sales		double	8	12
13	Profit		double	8	12
14	ValueSegment		varchar	10	10

Fig 2. Data Pelanggan TagMedia

B. *Exploratory Data Analysis*

Proses eksplorasi data dilakukan dalam dua tahapan, yaitu statistika deskriptif dan visualisasi.

a) *Descriptive statistic*

Proses statistika deskriptif dilakukan menggunakan SAS Studio untuk melihat nilai *mean*, *median*, *modus*, *variance*, *range*, *skewness*, *kurtosis*, *standard deviation*, dan lain-lain. Berikut merupakan hasil statistika deskrptif masing-masing variabel numerik,

Variable	Label	Mean	Std Dev	Minimum	Maximum	Median
Age	Age	26.8045198	12.8293349	12.0000000	60.0000000	23.0000000
Recency	Recency	242.4711864	312.5083791	0	1432.00	109.0000000
Frequency	Frequency	11.3966102	6.8897089	1.0000000	37.0000000	11.0000000
Sales	Sales	2619.05	2622.22	1.1880000	25043.05	1990.31
Profit	Profit	321.1519661	855.9761662	-6628.17	8979.92	192.1500000

Variable	Label	Variance	Range	Skewness	Kurtosis
Age	Age	164.5918348	48.0000000	0.9372946	-0.1893880
Recency	Recency	97661.49	1432.00	1.9705236	3.3444231
Frequency	Frequency	47.4680880	36.0000000	0.5498690	0.2516521
Sales	Sales	6876053.16	25041.86	2.4369149	10.4434144
Profit	Profit	732695.20	15608.09	1.9869111	27.6767700

Fig 3. Statistika Deskriptif Variabel Numerik

Sedangkan nilai distribusi dari variabel kategorikal adalah sebagai berikut. Berdasarkan distribusi ini, pelanggan paling banyak memberikan nilai *value* dalam tingkat *middle*. Pelanggan paling banyak berada merupakan tipe *consumer* (pelanggan biasa).

ValueSegment				
ValueSegment	Frequency	Percent	Cumulative Frequency	Cumulative Percent
High-Value	39	4.41	39	4.41
Low-Value	261	29.49	300	33.90
Mid-Value	585	66.10	885	100.00

CustomerType				
CustomerType	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Consumer	568	64.18	568	64.18
Corporate	166	18.76	734	82.94
Home Office	151	17.06	885	100.00

Fig 4. Statistika Deskriptif Variabel Kategorikal

b) *Visualization*

Eksplorasi data kedua dilakukan melalui visualisasi. Pertama, dilihat distribusi variabel numerik menggunakan histogram. Melalui histogram ini, diketahui bahwa distribusi data mayoritas tidak simetris. Pelanggan TagMedia didominasi oleh usia 10-25 tahun dengan keuntungan berada pada nilai 0 bahkan negatif. Hal ini menunjukan bahwa TagMedia masih perlu membenahi sistem pengambilan keuntungan dari penjualan produknya.

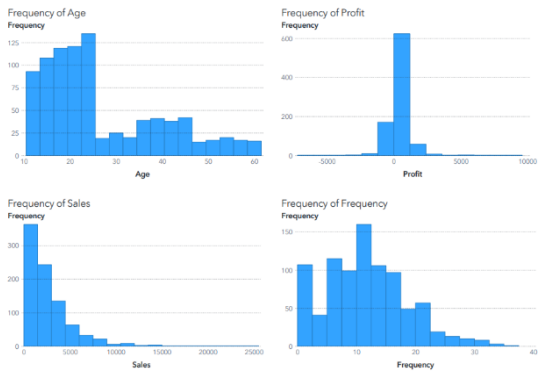


Fig 5. Distribusi Variabel Numerik

Eksplorasi selanjutnya dilakukan untuk mengetahui rata-rata *profit*, *recency*, *sales*, dan *frequency* pelanggan TagMedia berdasarkan *value segment*. Didapatkan bahwa pelanggan yang memiliki *low value* cenderung memiliki nilai *profit*, *sales*, dan frekuensi belanja yang rendah, namun nilai *recency* (jarak pembelian) sangat tinggi. Sebaliknya, pelanggan dengan *value* tinggi memberikan keuntungan, *sales*, dan frekuensi belanja yang tinggi. Jumlah rata-rata *recency* paling rendah diantara yang lain, hal ini berarti pelanggan sering berbelanja dengan rutin.

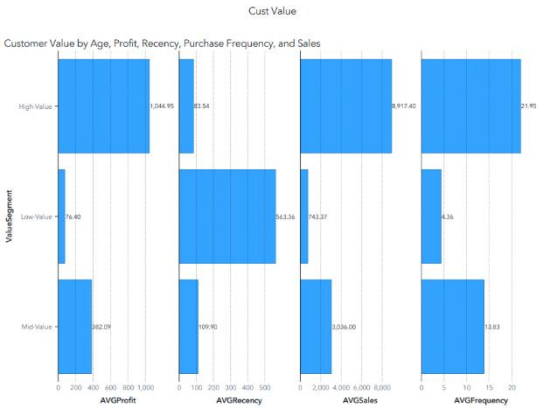


Fig 6. Rata-rata *Profit*, *Sales*, *Recency*, dan *Frequency* Pelanggan TagMedia Berdasarkan *Value Segment*

Diagram *boxplot* di bawah menunjukkan rentang nilai untuk setiap variabel numerik.

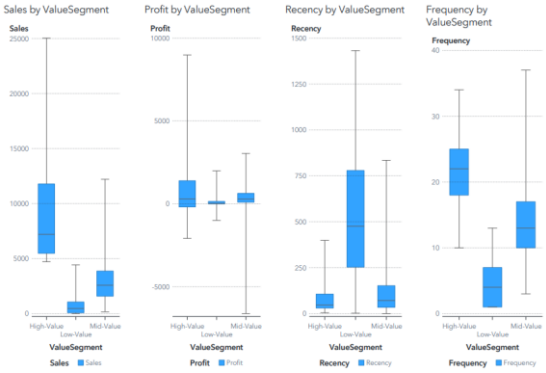


Fig 7. *Boxplot* Variabel Numerik

c) *Correlation*

Hubungan antarvariabel numerik digambarkan dalam *scatter plot* berikut ini. Warna grafik diambil dari *value segmentation*.

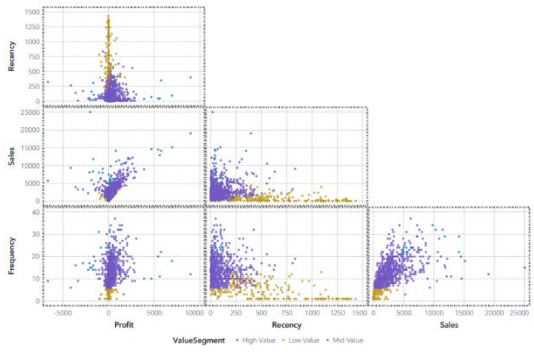


Fig 8. *Scatter Plot* Variabel Numerik

Secara lebih jelas, hubungan antarvariabel ditunjukkan melalui matriks korelasi Pearson di bawah. Hubungan bervariasi dari rentang rendah hingga sedang.

Pearson Correlation Coefficients, N = 885					
	Age	Recency	Frequency	Sales	Profit
Age Age	1.00000	0.02729	-0.00340	0.10233	-0.00977
Recency Recency	0.02729	1.00000	-0.52405	-0.30940	-0.13628
Frequency Frequency	-0.00340	-0.52405	1.00000	0.59611	0.24632
Sales Sales	0.10233	-0.30940	0.59611	1.00000	0.51467
Profit Profit	-0.00977	-0.13628	0.24632	0.51467	1.00000

Fig 9. Korelasi Variabel Numerik

C. *Modeling & Evaluation*

Data pelanggan kemudian dibentuk ke dalam model menggunakan algoritma yang berbeda. Variabel prediktor meliputi *sales*, *profit*, *frequency*, dan *recency*. Sedangkan variabel target adalah *value segment*. Data dibagi menjadi 70% *training* dan 30% *testing*.

a) *Decision Tree*

Hasil *decision tree* adalah sebagai berikut dengan nilai *validation misclassification rate* sebesar 0.0528. Model melakukan kesalahan prediksi pada data *testing* sebanyak 14 dari 264 data.

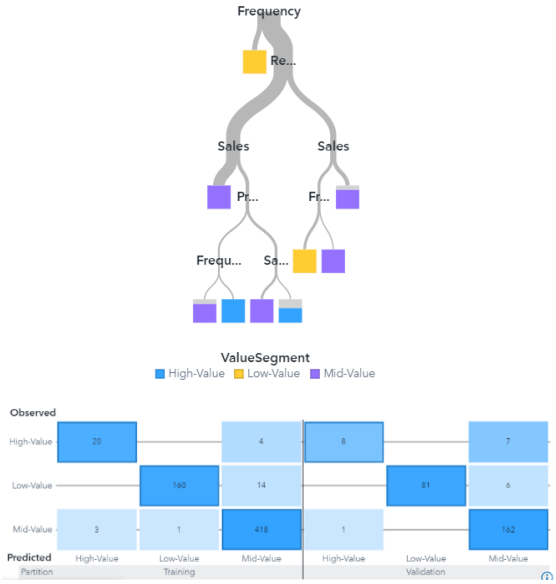


Fig 10. *Decision Tree Model and Evaluation*

b) *Neural Network*

Model *neural network* yang dibangun memiliki 10 neuron dengan hasil akhir 3 *value segment*. Hasil *misclassification rate* sebesar 0.0906 dengan jumlah kesalahan prediksi data *testing* sebesar 24 dari 264 data.

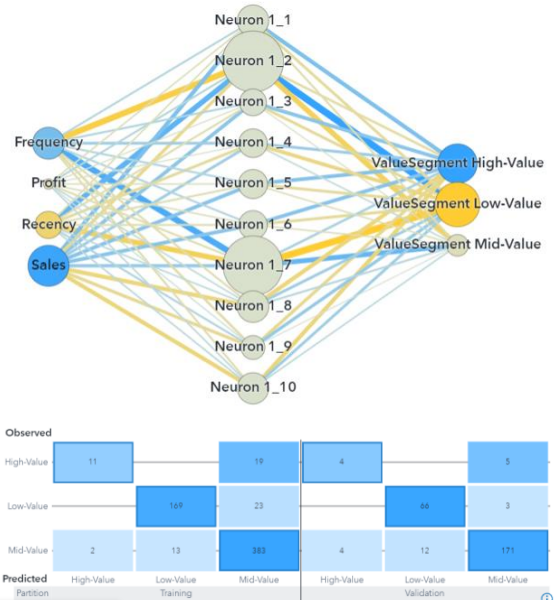


Fig 11. *Neural Network Model and Evaluation*

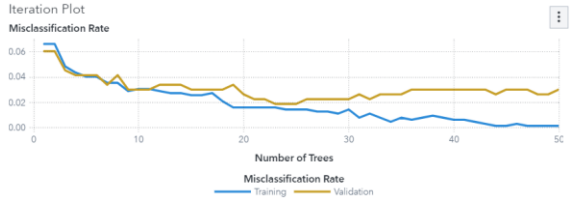
c) *Random Forest*

Model *Random Forest* yang dibangun membandingkan jumlah pohon yang berbeda-beda beserta dengan nilai *misclassification rate* untuk data *testing* dan *training*. Hasil *misclassification rate* sebesar 0.0415 dengan jumlah kesalahan prediksi data *testing* sebesar 11 dari 264 data.

Fig 12. *Random Forest Classifier Model and Evaluation*

d) *Gradient Boosting*

Model *Gradient Boosting* bekerja hampir sama dengan *Random Forest Classifier* yang membandingkan jumlah pohon klasifikasi mulai dari 0-50. Hasil kesalahan klasifikasi setiap jumlah pohon berbeda-beda. Namun, secara umum nilai *misclassification rate* untuk data *testing* sebesar 0.0302 dengan jumlah kesalahan prediksi data *testing* sebesar 8 dari 264 data.



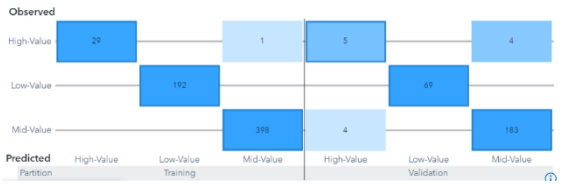


Fig 13. *Gradient Boosting Model and Evaluation*

D. *Model Selection*

Tahap terakhir dalam penelitian ini adalah membandingkan setiap model yang telah dibentuk menggunakan *Model Comparison* pada SAS Visual Analytics. Objek tersebut akan membandingkan setiap model yang telah dibuat dan memilih model paling optimal berdasarkan nilai *misclassification rate* terkecil. Secara umum, *Gradient Boosting* memiliki nilai kesalahan klasifikasi paling kecil, dengan demikian model ini dipilih sebagai model terbaik untuk kasus klasifikasi *value* pelanggan TagMedia.

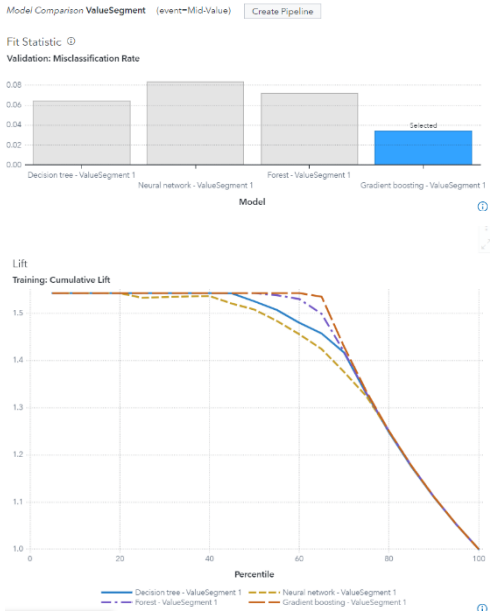


Fig 14. *Random Forest Classifier Model and Evaluation*

E. Diskusi

Proses *data science* pada kasus TagMedia dijalankan menggunakan tahapan dalam *framework* CRISP-DM menggunakan *software* SAS. Data pelanggan diolah menggunakan beberapa algoritma dengan hasil performa terbaik dimiliki oleh model *Gradient Boosting*. Model ini memiliki *misclassification* paling rendah diantara tiga algoritma lain, yaitu 0.0302. Berdasarkan hasil ini, TagMedia dapat menggunakan *Gradient Boosting* untuk melakukan prediksi *value* pelanggan. Hasil prediksi dan analisis tersebut dapat digunakan dan diterapkan untuk kegiatan pemasaran TagMedia. Berikut merupakan beberapa rekomendasi dari data:

- Pelanggan dalam kategori *low value* memiliki jarak belanja (dalam satuan hari) yang tinggi dan frekuensi belanja rendah, permasalahan ini mungkin menimbulkan *customer churn*. Dengan demikian, TagMedia harus memberikan promosi dan tawaran yang menarik agar pelanggan mulai berbelanja kembali. Pelanggan kategori ini umumnya memberikan *profit* yang rendah, dengan demikian TagMedia harus memperhitungkan kembali sistem pengambilan keuntungannya agar tidak ada produk yang menghasilkan *loss*.
- Pelanggan dalam kategori *middle value* memiliki jarak belanja sekitar 109 hari (3-4 bulan) dengan nilai *sales* dan profit yang cukup tinggi. Rata-rata transaksi sebanyak 13, transaksi ini masih berpotensi untuk meningkat. TagMedia perlu menyediakan layanan *customer service* yang baik dan cepat sehingga pelanggan kategori ini tidak mengurangi intensitas belanjanya di TagMedia.
- Pelanggan dalam kategori *high value* melakukan transaksi rutin dalam jangka waktu rata-rata 83 hari (2-3 bulan) dan memberikan *profit* besar bagi TagMedia. Pelanggan tipe ini harus diberikan pelayanan khusus dan ekslusif, sama seperti member tingkat *platinum* atau *diamond*. TagMedia juga bisa memberikan penawaran khusus yang hanya diberikan untuk pelanggan kategori *high value*.

V. KESIMPULAN DAN SARAN

A. *Kesimpulan*

TagMedia merupakan perusahaan *retail* yang menjual peralatan dan perlengkapan kantor. Dalam rangka meningkatkan performa bisnis, TagMedia perlu mengenali pelanggannya dengan lebih baik dan memperlakukan mereka secara khusus sesuai dengan perilaku belanjanya. Hal ini dapat diatasi dengan klasifikasi *customer value* untuk memprediksi pelanggan mana yang paling berharga, setia, dan menguntungkan terhadap TagMedia. Kategori *value* pelanggan perlu diberikan *treatment* khusus sesuai dengan kondisinya. Proses prediksi tipe *value* pelanggan mengikuti tahapan proses *data science* melalui *framework* CRISP-DM. Data pelanggan melalui serangkaian proses pengolahan dan pemodelan data sehingga menghasilkan *insight* serta model terbaik.

Pada proses *data science* ini, klasifikasi dilakukan menggunakan empat algoritma klasifikasi yang berbeda. Setiap performa model yang terbentuk dievaluasi dan dipilih yang terbaik. Performa tersebut diukur melalui *Misclassification Rate*. Hasil menunjukkan bahwa *Decision Tree* memiliki nilai 0.0528, *Neural Network* sebesar 0906, *Random Forest* sebesar 0.0415, dan *Gradient Boosting* sebesar 0.0302. Nilai *misclassification* diharapkan seminimal mungkin untuk mendapatkan model yang optimal. Berdasarkan hal tersebut, model terbaik berada pada *Gradient Bossting*. TagMedia dapat memanfaatkan model ini untuk memprediksi *customer value* dengan lebih akurat.

B. Saran

Berikut merupakan beberapa saran yang dapat diterapkan untuk melakukan penelitian dengan tema serupa:

a) Melakukan *feature engineering* lebih jauh untuk mendapatkan variabel-variabel lain yang dapat mendukung model.

b) Mencoba model kluster terlebih dahulu dengan cara yang berbeda sebelum melakukan klasifikasi.

c) Menggunakan algoritma klasifikasi lain sebagai pembanding sehingga mendapatkan model yang optimal.

REFERENCES

[1] Dhaliwal, G. S. (n.d.). Introduction to Retailing. Amity Directorate of Distance and Online Education.

[2] ETail Insights. (2021). How Many ETailers Are in The US? Retrieved from ETail Insights: https://www.etailinsights.com/online-retailer-market-size.

[3] Alsghaier, Hiba & Akour, Mohammed & Shehabat, Issa & Aldiabat, Samah. (2017). The Importance of Big Data Analytics in Business: A Case Study. American Journal of Software Engineering and Applications. Vol. 6. pp. 111-115. 10.11648/j.ajsea.20170604.12.

[4] Chen, H, Roger HL Chiang, and Veda C. Storey. "Business Intelligence and Analytics: From Big Data to Big Impact." MIS quarterly 36.4 (2012): 1165-1188.

[5] Jha, M., Jha, S., & O'Brien, L. (2016, June). Combining big data analytics with business process using reengineering. In Research Challenges in Information Science (RCIS), 2016 IEEE Tenth International Conference on (pp. 1-6). IEEE.

[6] Pratomo, E. A., Najib, M., & Mulyati, H. (2019). Customer Segmentation Analysis Based On The Customer Lifetime Value Method. *Journal of Applied Management, 17*(3), 408-415.

[7] Choudhury, Adil & Nur, Kamruddin. (2019). A Machine Learning Approach to Identify Potential Customer Based on Purchase Behavior. 10.1109/ICREST.2019.8644458.

[8] Favian, I. G., & Suryani, E. (2020). A Case Study of Applying Customer Segmentation in A Medical Equipment Industry. *International Conference on Management of Technology, Innovation, and Project* (pp. 119-125). Surabaya: Institut Teknologi Sepuluh Nopember.

[9] Schroer, C., Kruse, F., Gomez, & J. M. (2020). A Systematic Literature Review on Applying CRISP-DM Process Model. *International Conference on Enterprise Information System*.

[10] IBM Cloud Education. (2020, August 25). *Exploratory Data Analysis*. Retrieved from IBM: https://www.ibm.com/cloud/learn/exploratory-data-analysis.

[11] Maulidah, M., Gata, W., Aulianita, R., & Agustyaningrum, C. I. (2020). Algoritma Klasifikasi Decision TreeUntuk Rekomendasi Buku Berdasarkan Kategori Buku. *Jurnal Ilmiah Ekonomi dan Bisnis, 13*(2), 89-96.

[12] Farizawani, A. (2020). A review of artificial neural network learning rule based on multiple variant of conjugate gradient approaches. *Journal of Physics: Conference Series*. doi:10.1088/1742-6596/1529/2/022040.

[13] Gaurav. (2021, June 12). *An Introduction to Gradient Boosting Decision Trees*. Retrieved from Machine Learning Plus: https://www.machinelearningplus.com/machine-learning/an-introduction-to-gradient-boosting-decision-trees/.

[14] Yiu, T. (2019, June 12). *Understanding Random Forest*. Retrieved from Towards Data Science: https://towardsdatascience.com/understanding-random-forest-58381e0602d2.