Analisis Regresi Berganda terhadap Data E-Commerce

PATRICK ULYSSES
Program Studi Matematika
Universitas Katolik Parahyangan
Bandung, Indonesia
e-mail: 6162101009 @unpar.ac.id

***Abstract*—Dalam ranah e-commerce yang dinamis, memprediksi dan memahami Pengeluaran Tahunan (Yearly Amount Spent) pelanggan merupakan kebutuhan penting. Makalah ini memperkenalkan strategi inovatif berbasis data untuk memprediksi Pengeluaran Tahunan, dengan memanfaatkan dataset yang mencakup Average Session Length, Time on App, Time on Website, Length of Membership, dan Yearly Amount Spent. Dengan menerapkan teknik machine learning dan analisis regresi, penulis menyoroti pentingnya mengidentifikasi faktor-faktor paling berpengaruh dalam prediksi Pengeluaran Tahunan, yang pada gilirannya meningkatkan pemahaman perilaku pelanggan. Dengan mengeksplorasi berbagai model regresi dan mengevaluasi akurasi, interpretabilitas, dan kapasitas generalisasi mereka, kami memberikan alat yang kuat kepada bisnis untuk mengoptimalkan strategi pemasaran dan meningkatkan kinerja keuangan. Penelitian kami menekankan pentingnya segmentasi pelanggan, memungkinkan upaya pemasaran yang personal dan peningkatan retensi pelanggan, yang akan bermanfaat bagi para profesional e-commerce dan pengambil keputusan yang beroperasi di ranah e-commerce online yang dinamis.**

Multiple Regression, Data Analysis, Predictive Modelling, Machine Learning

I. PENDAHULUAN

Dalam lingkungan dunia digital yang terus berkembang berbelanja menggunakan platform e-commerce sudah menjadi hal yang *essential*. Kemampuan untuk memprediksi jumlah yang akan dibelanjakan pengguna pada platform akan menjadi elemen penting dan berharga bagi bisnis yang ingin mengoptimalkan strategi pemasaran, manajemen inventaris, dan meningkatkan pengalaman pelanggan. Untuk mengatasi hal ini, *multiple regression* telah muncul sebagai alat yang ampuh untuk membedah dan memahami berbagai faktor yang mempengaruhi keputusan pembelian pengguna.

Multiple regression adalah sebuah metode statistik yang memungkinkan kita untuk memeriksa hubungan kompleks antara berbagai variabel independen dan variabel dependen, dalam hal ini, jumlah uang yang akan dibelanjakan oleh seorang pengguna sebagai variable dependen. Platform e-commerce mengumpulkan beragam data untuk kemudian akan dimanfaatkan oleh penulis untuk menerapkan multiple regression, dalam hal ini kita dapat menyelidiki sejauh mana variabel yang telah dikumpulkan mempengaruhi perilaku belanja konsumen.

Pendekatan *multiple regression* tidak hanya memberikan wawasan mendalam mengenai perilaku konsumen namun juga memberdayakan platform e-commerce untuk membuat prediksi yang lebih akurat. Dengan memahami dinamika hubungan sebab-akibat yang mendasari keputusan pembelian, platform dapat menyesuaikan rekomendasi produk, strategi penetapan harga, dan promosi yang ditargetkan. *Multiple Regression* memungkinkan bisnis platform e-commerce meningkatkan pengalaman pelanggan dan efisiensi operasional.

Dalam makalah ini, penulis akan menggunakan konsep dasar *multiple regression* untuk menyoroti efektivitasnya dalam memprediksi pembelanjaan pengguna di platform e-commerce. Penulis juga akan mengeksplorasi data yang tersedia di platform e-commerce agar dapat dimanfaatkan untuk mengekstrak wawasan yang lebih dalam terhadap platform e-commerce yang bertujuan untuk meningkatkan kualitas platform.

II. METODE DAN DATA

Pendekatan dan metode yang penulis gunakan bertujuan untuk menganalisis dan memahami hubungan antara beberapa variabel dalam domain e-commerce. Sumber data yang penulis gunakan adalah dataset e-commerce yang tersedia di Kaggle [1], yang mencakup informasi tentang berbagai variabel yang relevan untuk penelitian ini. Tujuan utama penulis adalah untuk menyelidiki pengaruh variabel independen, yaitu Average Session Length, Time on App, Time on Website, dan Length of Membership, terhadap variabel dependen, yaitu Yearly Amount Spent.

No	Avg. Session Length	Time on App	Time on Website	Length of Membership	Yearly Amound Spent
1	34.49	12.65	39.577	4.082	587.95
2	31.92	11.10	37.26	2.66	392.20
3	33.00	11.33	37.11	4.10	487.54
...	...	...	...	...	...
499	33.32	12.39	36.84	2.33	456.46
500	33.71	12.41	35.77	2.73	497.77

Tabel 1. Data E-Commerce

Sebelum memasuki *multiple regression,* penulis akan melakukan analisis statistik ringkas yang mencakup perhitungan statistik deskriptif seperti rata-rata, median, standar deviasi, serta visualisasi data dalam bentuk histogram, box plots, dan scatter plots. Analisis ini memberi penulis wawasan awal tentang distribusi data dan hubungan antar variabel, yang akan membantu dalam interpretasi hasil dari *multiple regression* yang akan datang.

Dalam makalah ini, kami memilih menggunakan multiple regression sebagai metode statistik utama untuk mengidentifikasi pengaruh variabel independen terhadap variabel dependen. Regresi berganda memungkinkan kami untuk mengukur dan mengestimasi hubungan antara beberapa variabel independen, yaitu Average Session Length, Time on App, Time on Website, dan Length of Membership, dengan variabel dependen, yaitu Yearly Amount Spent. Dengan pendekatan ini, kami dapat memahami sejauh mana masing-masing variabel independen berkontribusi terhadap variasi dalam Yearly Amount Spent.

Rumus model regresi berganda yang akan kami gunakan untuk analisis ini adalah sebagai berikut:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

Dalam rumus di atas, Yearly Amount Spent (Y) adalah variabel dependen yang mewakili jumlah pembelian tahunan pelanggan. β0 adalah konstanta regresi, sementara β1, β2, β3, dan β4 adalah koefisien regresi yang akan kita estimasi. Variabel independen, yaitu *Average Session Length, Time on App, Time on Website,* dan *Length of Membership,* masing-masing memiliki koefisien yang menggambarkan pengaruh mereka terhadap Yearly Amount Spent. ε adalah kesalahan residual yang mencerminkan variabilitas yang tidak dapat dijelaskan oleh variabel independen dalam model ini. Melalui *multiple regression* ini, menggunakan bantuan *Python* penulis akan mengestimasi nilai koefisien β dan mengevaluasi signifikansi statistik dengan uji ANOVA serta menginterpretasikan pengaruh masing-masing variabel independen terhadap variabel dependen *Yearly Amount Spent*. Analisis ini akan memberikan pemahaman yang lebih mendalam tentang faktor-faktor yang memengaruhi perilaku pembelian pelanggan dalam konteks e-commerce.

III. Analisis dan Hasil

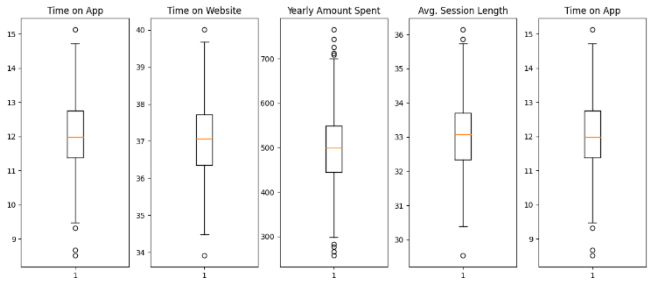
Pertama, menggunakan bantuan software python akan ditampilkan *summary statistics*. Berikut merupakan luaran *summary statistics* yang dihasilkan oleh python.

	Avg. Session Length	Time on App	Time on Website	Length of Membership	Yearly Amount Spent
Mean	33.05	12.05	37.06	3.53	499.31
Std	0.99	0.99	1.01	0.99	79.31
Min	29.53	8.50	33.91	0.26	256.67
Max	36.13	15.12	40.00	6.92	765.51
25 %	32.34	11.38	36.34	2.93	445.03
50 %	33.08	11.98	37.06	3.53	498.88
75%	33.71	12.75	37.71	4.12	549.31

Tabel 2 : Summary Statistics

Dalam analisis dataset ini, kami akan mengeksplorasi *variabel Average Session Length, Time on App, Time on Website, Length of Membership,* dan *Yearly Amount Spent.* Eksplorasi statistik ini memberikan wawasan yang cukup menarik tentang perilaku pelanggan dalam platform e-commerce. Secara rata-rata, pelanggan menghabiskan sekitar 33.05 menit dalam satu sesi, dengan pembagian waktu sebanyak 12.05 menit di aplikasi dan 37.06 menit di situs web. Rata-rata *membership* mencapai 3.53 tahun, sedangkan pengeluaran tahunan mencapai sekitar $499.31. Deviasi standar yang relatif kecil menunjukkan bahwa variasi dalam dataset terbatas, menunjukkan bahwa data cenderung berkelompok di sekitar nilai rata-rata masing-masing. Nilai minimum mengungkapkan data paling rendah untuk setiap variabel, mengungkapkan *Average Session Length* terendah di 29.53 menit, *Time on App* terpendek di 8.50 menit, *Time on Website* terendah di 33.91 menit, *Length of Membership* paling singkat hanya 0.26 tahun, dan *Yearly Amount Spent* terendah, yang mencapai $256.67. Sebaliknya, nilai maksimum menunjukkan titik tertinggi dalam dataset, dengan Average Session Length terpanjang di 36.13 menit, Time on App terpanjang di 15.12 menit, Time on Website terpanjang di 40.00 menit, masa keanggotaan mencapai puncak 6.92 tahun, dan Yearly Amount Spent yang paling signifikan, mencapai $765.51. Nilai kuartil, terutama nilai median (50%) memberikan pemahaman lebih mendalam tentang distribusi data.Nilai median setiap variabel yang cenderng serupa dengan rata-rata mengindikasikan distribusi data yang relatif simetris, di mana sekitar setengah dari observasi berada di atas dan di bawah titik pusat ini.

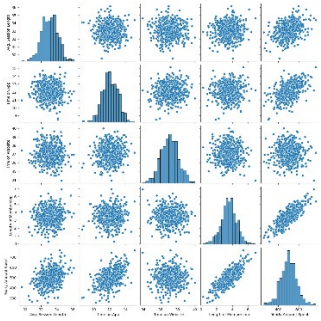
Berdasarkan hasil sebelumnya diperoleh bahwa seluruh variabel cenderung memiliki distribusi yang simetris untuk membuktikan hal tersebut akan digambarkan box plot dari setiap variabel.



Gambar 1. Boxplot variabel numerik

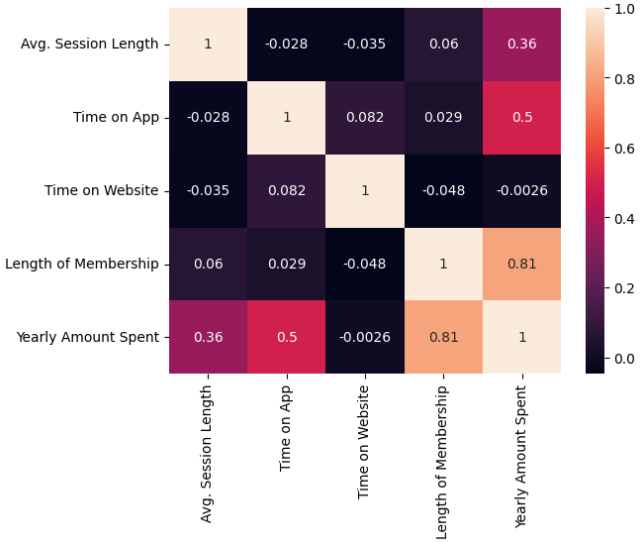
Berdasarkan gambar di atas maka dapat disimpulkan bahwa hasil observasi yang telah dilakukan sebelumnya benar, yaitu mayoritas variabel memiliki distribusi yang cenderung simetris dengan kecenderungan sentral, hal ini ditandai dengan *box* yang cenderung terletak pada posisi tengah dan sebaran garis dan *whiskers* yang simetris antar atas dan bawah. Hal ini berarti sebagian besar data berkumpul di sekitar nilai median dan juga nilai rata-rata dengan sebaran yang cukup terpusat. Selain itu terdapat juga beberapa data outliers pada setiap bariabel dengan data *Yearly Amount Spent* hal ini menandakan bahwa variabilitas data *Yearly Amount Spent* paling besar jika dibandingkan dengan variabel lainnya.

Selanjutnnya akan diselidiki hubungan antar variabel-variabel dalam data dengan menggunakan matriks plot scatter. Adapun luaran yang dihasilkan seperti berikut.



Gambar 2. Hasil *scatter plot* dan sebaran variabel numerik

Berdasarkan hasil di atas, diperoleh bahwa mayoritas variabel memiliki hubungan yang lemah terhadap satu sama lain, hal ini ditandai dengan plot scatter mayoritas variabel yang cenderung tersebar dan tidak berpola. Namun amati bahwa variabel dependen *Yearly Amount Spent* memiliki relasi linear yang sangat kuat terhadap variabel *Length of Membership* (Terlihat dari sebaran scatter plot yang dapat direpresentasikan terhadap garis), relasi linear positif yang cukup kuat terhadap *Time on App* (Terlihat dari sebaran scatter plot yang dapat cukup baik direpresentasikan oleh garis lurus), dan memiliki relasi linear positif yang cukup baik terhadap variabel *Avg Session Length* (Terlihat dari sebaran scatter plot yang dapat dengan cukup direpresentasikan garis lurus). Untuk membuktikan hal di atas selanjutnya akan digunakan analisis matriks korelasi untuk mengukur seberapa kuat ataupun lemah pengaruh hubungan linear antar variabel dalam data set. Berikut merupakan luaran hasil matriks korelasi.



Gambar 3. Heatmap Korelasi

Sebelum memasuki analisis, perlu diingat bahwa nilai korelasi memiliki rentang antara -1 hingga 1, di mana nilai -1 menandakan relasi linear negatif yang kuat sedangkan 1 menandakan relasi linear positif yang kuat. Amati bahwa pada matriks di atas, sesuai hasil yang diperoleh pada bagian sebelumnya variabel dependen *Yearly Amount Spent* memiliki relasi linear yang sangat kuat terhadap variabel *Length of Membership,* relasi linear positif yang cukup kuat terhadap *Time on App* dan memiliki relasi linear positif yang terhadap variabel *Avg Session Length.* Hal ini menandakan bahwa peningkatan terhadap variabel *Length of Membership, Time on App,* dan *Avg Session Length* akan meningkatkan nilai *Yearly Amound Spent*. Oleh karena itulah, peningkatan variabel *Length of Membership, Time on App,* dan *Avg Session Length* menjadi kunci untuk meningkatkan *Yearly Amound Spent*. Berdasarkan hasil di atas, penulis juga menduga model konstanta regresi untuk variabel *Length of Membership, Time on App,* dan *Avg Session Length* akan lebih signifikan jika dibanding dengan 2 variabel lainnya.

Selanjutnya akan diterapkan *multiple regression* terhadap data set e-commerce untuk memprediksi nilai *Yearly Amound Spent.* Menggunakan bantuan python diperoleh bahwa konstanta regresi adalah sebagai berikut.

$$Y = -1.037 \text{ x } 10^{3} + 2.576 \text{ x } 10\ X_1 + \ 3.8801 \text{ x } 10\ X_2 + \\ -1.804 \text{ x } 10^{-2}\ X_3 + 6.185\ x\ 10\ X_4 + \ \varepsilon$$

Di mana :

- X1 menyatakan *Avg Session Length*
- X2 menyatakan *Time on App*
- X3 menyatakan Time on Website
- X4 menyatakan Length of Membership
- ε menyatakan error yang disebabkan oleh variabel lainnya yang tidak ada di data set

Perhatikan bahwa berdasarkan hasil di atas, variabel *Avg Session Length, Time on App,* dan Length of *Membership* memiliki pengaruh yang positif terhadap variabel dependen *Yearly Amount Spent*, hal tersebut terlihat dari koefesien regresi yang bernilai positif lebih tepatnya kenaikan *Time on App* sebesar 1 menit akan meningkatkan *Avg Session Length* sebanyak 2.576, kenaikan *Time on App* sebesar 1 menit akan menaikan *Yearly Amount Spent* sebanyak 3.8801, dan kenaikan *Lenth of Membership* sebanyak 1 tahun akan meningkatkan *Yearly Amount Spent* sebanyak 6.185. Sedangkan variabel *Time on Website* memiliki pengaruh yang negatif terhadap *Yearly Amount Spent* yang berarti kenaikan *Time on Website* sebesar 1 menit akan menurunkan *Yearly Amount Spent* sebesar $-1.804 \text{ x } 10^{-2}$. Hasil di atas sesuai dengan dugaan yang sebelumnya di mana variabel yang akan berpengaruh signifikan terhadap variabel dependen adalah variabel yang memiliki nilai korelasi yang kuat terhadap variable dependen, dalam konteks ini variabel *Length of Membership, Time on App,* dan *Avg Session Length*. Selanjutnya akan diukur signifikansi dari konstanta *multiple regression* yang telah dibuat menggunakan uji ANOVA. Diperoleh tabel ANOVA sebagai berikut

No.	SS	Df	Ms	Fs	pval
0	949765.3	5	189953	1928.5	1.1 x 10^{-16}
1	14183.6	144	98.49	...	
2	940952.6	149	...	...	

Tabel 3. Tabel ANOVA

Berdasarkan hasil di atas dapat disimpulkan bahwa konstanta regresi yang telah didapat signifikan hal ini ditandai dengan p-value yang bernilai rendah, yaitu 1.1 x 10^{-16}. Hal tersebut berarti variabel di atas memiliki berpengaruh terhadap variabel dependen *Yearly Amound Spent* di mana berdasarkan hasil sebelumnya telah diketahui bahwa kenaikan *Time on App* sebesar 1 menit akan meningkatkan *Avg Session Length* sebanyak 2.576, kenaikan *Time on App*

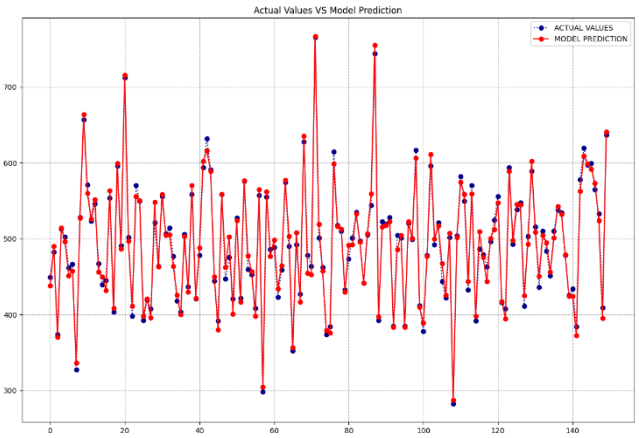
sebesar 1 menit akan menaikan *Yearly Amount Spent* sebanyak 3.8801, dan kenaikan *Lenth of Membership* sebanyak 1 tahun akan meningkatkan *Yearly Amount Spent* sebanyak 6.185. Hal di atas berarti kunci dari peningkatan *Yearly Amount Spent* terletak pada variabel *Length of Membership, Time on App,* dan *Avg Session Length* dengan variabel yang paling signifikan adalah *Length of Membership* karena memiliki kenaikan *Length of Membership* akan meningkatkan *Yearly Amount Spent* lebih besar jika dibanding dengan variabel lainnya peningkatan *Length of Membership* bisa dilakukan dengan memberikan promo-promo ataupun diskon menarik terhadap yang berlangganan, namun perlu diingat bahwa meningkatkan *Length of Membership* adalah hal yang cukup sulit dilakukan dan memerlukan sumber daya yang sangat banyak sehingga fokus terbesar harus dikerahkan ke peningkatan variabel *Time on App,* dan *Avg Session Length* karena walaupun dampaknya terhadap *Yearly Amount Spent* tidak sebesar *Length of Membership* namun untuk meningkatkan variabel *Time on App,* dan *Avg Session Length* akan jauh lebih mudah, peningkatan variabel *Time on App* dan *Avg Session Length* bisa dilakukan dengan melakukan peningkatan terhadap app dan web sehingga lebih interaktif dan menarik agar pengguna semakin senang menggunakan platform e-commerce. Selain itu variabel *Time on Web* memiliki dampak yang negatif terhadap variabel *Yearly Amount Spent* hal ini berarti bahwa website platform e-commerce harus dioptimasi agar dapat memberikan dampak yang positif terhadap variabel *Yearly Amount Spent*.

Selanjutnya kami akan menilai performa model yang telah dibuat. Menggunakan MSE dan R^2 akan diukur peforma model regresi yang telah dibuat. Perhatikan tabel di bawah.

Nilai R^2	Nilai MSE
0.9849262667370814	91.55779479261258

Tabel 4. Nilai R^2 dan MSE

Berdasarkan hasil di atas, model *multiple regression* yang telah dibuat mempunyai performa yang sangat baik yang ditandai dengan nilai R^2 yang tinggi yaitu, 0.98. Hal ini berarti bahwa model mampu memprediksi sesuai dengan data yang ada, mampu menangkap dan menjelaskan sebagian besar variasi yang diamati dalam variabel dependen. Nilai R-squared yang tinggi juga sering menunjukkan hubungan yang kuat dan jelas antara variabel independen dan dependen. Semakin tinggi nilai R-squared, semakin kuat hubungannya. Nilai R-squared sebesar 0,98 berarto bahwa model ini kemungkinan besar mampu memberikan prediksi yang akurat. Ini menunjukkan bahwa model dapat efektif membuat prediksi berdasarkan variabel independen. Terakhir, sekitar 98% dari variabilitas dalam variabel dependen dapat diatribusikan kepada variabel independen yang dimasukkan dalam model ini. Ini berarti bahwa ada sedikit variabilitas yang tidak dapat dijelaskan atau acak yang tersisa dalam data. Selain itu, nilai MSE (*Mean Square Error)* dari model di atas juga cukup rendah, yaitu sekitar 91.55. MSE yang rendah menunjukkan bahwa prediksi model mendekati dengan baik data aktual. Ini berarti bahwa model mampu mengestimasi variabel dependen dengan baik berdasarkan variabel independen. Selain itu, MSE yang rendah menandakan bahwa prediksi model akurat dan memiliki variasi yang rendah, artinya kesalahan prediksi berada dalam jangkauan yang terbatas di sekitar nilai-nilai aktual. Nilai MSE yang rendah tersebut merupakan tanda bahwa model yang cocok dan dapat menjelaskan data dengan baik, efektif dalam menangkap tren dan pola,serta hubungan mendasar antar variabel dalam data tersebut. Ini mengindikasikan bahwa model tersebut memiliki kualitas yang tinggi dalam hal akurasi prediksi dan merupakan indikasi bahwa model tersebut cocok dengan baik untuk data, meminimalkan kesalahan prediksi dan memberikan perkiraan yang dapat diandalkan. Untuk menunjukan akurasi dan presisi model perhatikan gambar di bawah.



Gambar 4. Hasil model terhadap data testing

Gambar di atas merupakan perbandingan nilai aktual data testing terhadap data prediksi hasil model *multiple regression* dapat dilihat berdasarkan hasil di atas bahwa model *multiple regression* untuk data e-commerce dapat dengan sangat baik memprediksi variabel dependen *Yearly Amount Spent* dengan selisih dengan nilai aktual yang cukup rendah.

IV. KESIMPULAN

Berdasarkan hasil yang telah dibahas pada bagian sebelumnya dapat disimpulkan bahwa

1. Mayoritas variabel numerik dalam data set e-commerce memiliki distribusi yang cenderung terpusat terhadap nilai median dan mean dengan variabilitas yang rendah
2. Mayoritas variabel numerik dalam data set e-commerce memiliki hubungan linear yang sangat lemah. Namun variabel *Yearly Amount Spent* memiliki relasi linear yang sangat kuat terhadap variabel *Length of Membership,* relasi linear positif yang cukup kuat terhadap *Time on App* dan memiliki relasi linear positif yang terhadap variabel *Avg Session Length.*
3. Variabel *Length of Membership* memiliki pengaruh yang paling signifikan namun paling sulit untuk ditingkatkan sedangkan variabel *Time on App* dan *Avg Session Length* memiliki dampak yang positif namun tidak sebesar *Length of Membership.* Namun mengingat bahwa meningkatkan *Time on App* dan *Avg Session Length* jauh lebih mudah dan memerlukan sumber daya yang lebih sedikit maka

peningkatan *Time on App* dan *Avg Session Length* harus diutamakan

4. Variabel *Time on Web* memiliki dampak yang negatif terhadap variabel dependen. Oleh karena itu, website e-commerce harus dioptimasi kembali agar dapat memberikan dampak yang positif terhadap variabel dependen.
5. Performa model *multiple regression* yang digunakan memiliki performa yang sangat baik dan dapat memprediksi nilai *Yearly Amount Spent* dengan sangat baik yang terindikasi melalui uji ANOVA yang signifikan, nilai MSE yang rendah, dan nilai R^2 yang tinggi.

REFERENCES

[1]. *Linear Regression E-commerce Dataset.* (2019, September 16). Kaggle. https://www.kaggle.com/datasets/kolawale/focusing-on-mobile-app-or-website

[2]. *What is E-commerce in Indonesia : Definition, Types, and Benefits.* (n. d.). https://developers.bri.co.id/end/node/50787

[3]. Junianti, S., Yulita, H., & Christian, M. (n.d.). A MULTIPLE REGRESSION ANALYSIS OF TOKOPEDIA E-COMMERCE USERS' PURCHASING DECISIONS. ResearchGate. https://www.researchgate.net/publication/368307072_A_MULTIPLE_REGRESSION_ANALYSIS_OF_TOKOPEDIA_E-COMMERCE_USERS'_PURCHASING_DECISIONS

[4]. Team, S. O. A. (n.d.). What Is Ecommerce? Definition, Types, Advantages, and Disadvantages. Amazon. https://sell.amazon.com/learn/what-is-ecommerce#:~:text=Ecommerce%20is%20a%20method%20of,media%20to%20drive%20online%20sales.