Pendekatan Data Science untuk Menemukan Churn Pelanggan pada Sector Perbankan dengan Machine Learning

Amir Mahmud Husein[1*], Mawaddah Harahap[2,] Piter Fernandito[3]
[1,2]Univeritas Prima Indonesia, Fakultas Teknologi dan Ilmu Komputer, Teknik Informatika, Indonesia,
[1]amirmahmud@unprimdn.ac.id, [2]mawaddah@unprimdn.ac.id, [3]piterfernandito@gmail.com

Received: 1 Mey 2021 Accepted: 9 May 2021 Published: 14 May 2021

***Amir Mahmud Husein**

Keywords: Data Science, Analisis Regresi, Churn, Banking, Prediksi Churn Customer

Abstrak : Peralihan pelanggan merupakan fenomena dimana pelanggan perusahaan berhenti membeli atau berinteraksi sehingga sangat penting bagi perusahaan khususnya perbankan untuk memprediksi kemungkinan churn pelanggan dan hasilnya dapat digunakan untuk membantu retensi pelanggan dan bagian dari strategi perusahaan. Makalah ini menyajikan analisis dan prediksi churn pelanggan dengan menggunakan lima model berbeda yaitu Kneighbors Classifier, Logistic Regression, Linear SVC, Random Tree Classifier dan Random Forest Classifier. Berdasarkan hasil pengujian pendekatan model Random Forest Classifier dan Kneighbors Classifier lebih baik dari pada model lain dengan akurasi sebesar 86% dan 84%. Rekayasa fitur dengan pendekatan Anova dan Chi Square memiliki pengaruh yang signifikan terhadap peningkatan kinerja model prediksi.

PENDAHULUAN

Peralihan pelanggan merupakan fenomena dimana pelanggan perusahaan berhenti membeli atau berinteraksi dengan perusahaan. Churn rate yang tinggi berarti banyak pelanggan yang tidak mau membeli barang atau jasa dari perusahaan (Ahmed and Maheswari 2019). Churn pelanggan adalah perhitungan matematis dari persentase pelanggan yang tidak mungkin melakukan pembelian lagi dari perusahaan. Menghindari gesekan pelanggan adalah salah satu faktor kunci dalam memaksimalkan profitabilitas organisasi seperti bank (Keramati, Ghaneei, and Mirmohammadi 2016; Sayed, Abdel-Fattah, and Kholief 2018; Bilal Zoric 2016; Kaur and Kaur 2020; Karvana et al. 2019) dan layanan telekomunikasi (Zhao et al. 2021; Ahmed and Maheswari 2019; Ahmad, Jafar, and Aljoumaa 2019). Di sektor perbankan, inovasi teknologi saat ini memberikan kemudahan bagi pelanggan untuk membuka rekening bank baru dan mentransfer semua aset bahkan tanpa harus datang dimana pelanggan dapat melakukannya dari rumah. Situasi ini membuat bank lebih tertarik pada topik loyalitas pelanggan. Namun, sebelum memulai cara yang efisien untuk mempertahankan pelanggan yang sudah ada, sangat perlu memprediksi pelanggan mana yang akan berhenti (Szmydt 2019).

Mengakhiri hubungan pelanggan dengan perusahaan memiliki nilai yang tidak dapat disangkal untuk semua organisasi karena memungkinkan mereka untuk menyiapkan kampanye bertarget untuk mempromosikan loyalitas pelanggan (Szmydt 2019). Prakiraan churn yang akurat secara efektif mendukung strategi loyalitas pelanggan dan merencanakan kampanye pemasaran ekonomi, menghasilkan penghematan yang signifikan bagi penyedia layanan. Untuk bertahan dari persaingan yang ketat ini, perusahaan telekomunikasi menjadi lebih agresif dengan berinvestasi lebih banyak dalam mengembangkan data mining dan model berbasis pembelajaran mesin untuk analitik, peramalan, dan manajemen churn (Zhao et al. 2021). anyak penelitian telah menemukan bahwa teknologi pembelajaran mesin sangat efisien dalam memprediksi situasi ini. Berbagai pendekatan pembelajaran mesin untuk prediksi churn telah diusulkan dalam literatur. Namun, tugas ini sangat sulit karena distribusi kelas yang tidak merata dan jumlah kelas pelanggan yang tidak

dibatalkan melebihi jumlah kelas churn (Ahmad, Jafar, and Aljoumaa 2019). Belajar dari distribusi kelas yang tidak seimbang sangat kompleks dengan sebagian besar algoritma pembelajaran mesin klasik, karena cenderung mengklasifikasikan sebagian besar kelas dengan benar dan mengabaikan yang langka. Masalah besar lainnya adalah ada banyak faktor yang mempengaruhi gesekan pelanggan dan hubungan di antara mereka sangat kompleks (Sabbeh 2018). Mengukur peran faktor-faktor ini juga merupakan tugas yang kompleks (Sabbeh 2018). Oleh karena itu, interpretasi model pembelajaran mesin sangat penting. Sebagian besar studi literatur sebelumnya berfokus pada peningkatan akurasi model dan kurang tertarik untuk memahami model yang dihasilkan dan mengukur peran faktor-faktor yang mempengaruhi keakuratan klasifikasi.

Makalah ini membahas tentang penerapan data science dalam menyajikan analisis dan menemukan menemukan chun pelanggan pada sektor perbankan. Tujuan utama dari makalah ini adalah:

- Menganalisis dan mencari keadaan churn pelanggan
- Melakukan segmentasi pelanggan
- Memprediksi jumlah persentase nasabah yang keluar dari bank dengan analisis regresi
- Menemukan pelanggan yang memiliki potensi akan meninggalkan bank

TINJAUAN LITERATUR

Banyak pendekatan yang diterapkan untuk memprediksi churn di perusahaan. Sebagian besar pendekatan ini telah menggunakan pembelajaran mesin dan penambangan data. Sebagian besar pekerjaan terkait berfokus pada penerapan hanya satu metode penambangan data untuk mengekstrak pengetahuan, dan yang lainnya berfokus pada membandingkan beberapa strategi untuk memprediksi churn. Kepuasan pelanggan dan churn pelanggan telah diidentifikasi sebagai dua faktor yang berkontribusi terhadap keberhasilan industri dan oleh karena itu topik hangat diteliti di berbagai industri, seperti pariwisata (Hassan et al. 2015), telekomunikasi (Ahmed and Maheswari 2019) dan banking (Ashish et al. 2021; Bilal Zoric 2016).

Ahmad et al (Ahmad, Jafar, and Aljoumaa 2019) mengembangkan model perkiraan churn yang membantu operator memprediksi pelanggan mana yang kemungkinan besar akan churn. Model yang dikembangkan dalam makalah ini menggunakan teknologi pembelajaran mesin pada platform data besar dan dibangun di atas metode rekayasa dan pemilihan fitur baru. Pengukuran under-standard curve area (AUC) digunakan untuk mengukur kinerja model, dan diperoleh nilai AUC sebesar 93,3%. Kontribusi penting lainnya dalam makalah ini adalah penggunaan jaringan pelanggan sosial dalam model prediktif dengan mengekstraksi fungsi analisis jaringan sosial (SNA).

Ammar et al (Ahmed and Maheswari 2019) mengusulkan strategi berbasis baru untuk memprediksi tingkat churn perusahaan telekomunikasi. model prediksi churn ensemble yang digabungkan dengan pemodelan uplift berbasis biaya. Model ini menguntungkan organisasi dengan memenuhi tujuan bisnis secara efektif dan juga dengan mengurangi biaya secara signifikan jika dibandingkan dengan model lainnya. Pendekatan ensemble yang diusulkan meningkatkan proses prediksi dan memberikan akurasi 94,3%.

Almuqren et al (Almuqren, Alrayes, and Cristea 2021) mengusulkan pendekatan baru dengan menggunakan penambangan media sosial untuk memprediksi churn pelanggan di bidang telekomunikasi. Metode baru yang diusulkan digunakan untuk mengekstrak umpan balik kepuasan pelanggan secara real-time dan digunakan untuk memprediksi churn pelanggan dengan tingkat akurasi 95,8%.

Kaur, Ishpreet, Kaur, Jasleen (Kaur and Kaur 2020) menerapkan model pembelajaran mesin yang berbeda seperti *Logistic regression* (LR), *decision tree* (DT), *K-nearest neighbor* (KNN), *random forest* (RF) dan lainnya diterapkan pada dataset bank untuk memprediksi probabilitas pelanggan yang akan melakukan churn. Mereka malaporkan Random Forest berkinerja lebih baik di antara semua model yang digunakan.

Faris, H (Faris 2018) mengusulkan model hybrid cerdas berdasarkan Particle Swarm Optimization dan jaringan saraf Feedforward diusulkan untuk prediksi churn. PSO digunakan untuk menyesuaikan bobot fitur input dan mengoptimalkan struktur jaringan saraf secara bersamaan untuk meningkatkan daya prediksi dan (Semrl and

Matei 2017) menerapkan memprediksi perilaku pelanggan, untuk meningkatkan pemanfaatan gym dan retensi pelanggan.

BAHAN DAN METODE

Bahan

Data yang digunakan dalam penelitian ini merupakan kumpulan data churn banking bersumber dari kaggel https://www.kaggle.com/santoshd3/bank-customers. Dataset ini memiliki attribut RowNumber, CustomerId, Surname, CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary, Exited. Dari atribut ini, CreditScore, Age, Tenure, Balance, NumOfProducts, EstimatedSalary adalah variabel diskrit dan sisanya adalah variabel kategoris. Pada tabel 1 berikut ini sebagian data yang digunakan.

Tabel 1 Himpunan data

RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1	15634602	Hargrave	619	France	Female	42	2	0	1	1	1	101348.88	1
2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
3	15619304	Onio	502	France	Female	42	8	159660.8	3	1	0	113931.57	1
4	15701354	Boni	699	France	Female	39	1	0	2	0	0	93826.63	0
5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.1	0
6	15574012	Chu	645	Spain	Male	44	8	113755.78	2	1	0	149756.71	1
7	15592531	Bartlett	822	France	Male	50	7	0	2	1	1	10062.8	0
8	15656148	Obinna	376	Germany	Female	29	4	115046.74	4	1	0	119346.88	1
9	15792365	He	501	France	Male	44	4	142051.07	2	0	1	74940.5	0
10	15592389	H?	684	France	Male	27	2	134603.88	1	1	1	71725.73	0
11	15767821	Bearce	528	France	Male	31	6	102016.72	2	0	0	80181.12	0
12	15737173	Andrews	497	Spain	Male	24	3	0	2	1	0	76390.01	0
13	15632264	Kay	476	France	Female	34	10	0	2	1	0	26260.98	0
14	15691483	Chin	549	France	Female	25	5	0	2	0	0	190857.79	0
15	15600882	Scott	635	Spain	Female	35	7	0	2	1	1	65951.65	0
16	15643966	Goforth	616	Germany	Male	45	3	143129.41	2	0	1	64327.26	0
17	15737452	Romeo	653	Germany	Male	58	1	132602.88	1	1	0	5097.67	1
18	15788218	Henderson	549	Spain	Female	24	9	0	2	1	1	14406.41	0
19	15661507	Muldrow	587	Spain	Male	45	6	0	1	0	0	158684.81	0
20	15568982	Hao	726	France	Female	24	6	0	2	1	1	54724.03	0

Metode

Pada gambar 1 merupakan ilustrasi metodologi yang digunakan pada penelitian ini. Secara umum terdapat 5 langkah yang digunakan untuk menemukan dan memprediksi churn pelanggan pada himpunan data yaitu pengumpulan data dan explorasi untuk lebih memahami data. Pembersihan data digunakan untuk memastikan kumpulan data tidak ada yang duplikat, data yang hilang, outler dan menentukan fitur data yang memiliki arti penting untuk keperluan analisis dan prediksi. Selanjutnya ekstraksi fitur merupakan salah satu pendekatan yang digunakan untuk merekayasa fitur penting dalam model. Pada tahapan ini analisis ANOVA diusulkan untuk mengevaluasi korelasi antar fitur dan terkahir digunakan model Chi Square untuk memilih fitur yang akan digunakan pada model. Tahapan pre-processing digunakan untuk memisahkan data latih dan data uji dan akhirnya model regresi digunakan untuk memprediksi chun pelanggan.

Gambar 1 Metodologi yang diusulkan

HASIL PENELITIAN

Pada bagian ini akan kami uraikan hasil prediksi churn pelanggan. Setelah melakukan proses pengumpulan data, explorasi dan pembersihan data, salah satu tahapan penting sebelum menerapkan model regresi adalah ekstraksi fitur. Ekstraksi fitur adalah proses mengambil kumpulan data dan membuat variabel penjelas, atau

fitur prediktor, yang kemudian diteruskan ke model prediksi untuk melatih algoritme pembelajaran mesin. Dalam hal ini, kami menggunakan dua pendekatan yaitu analisis ANOVA dan Chi Square. Berdasarkan pengamatan dari ANOVA, kami menemukan bahwa CreditScore, Age, Balance, Tenure dan NumOfProduct adalah fitur penting. Pengamatan dengan Chi Square, kami menemukan bahwa Geografi, Gender dan IsActiveMember adalah fitur penting. Dari hasil keduanya, kami menyimpulkan CreditScore, Age, Balance, NumOfProduct, Geografi, Gender, Tenure dan IsActiveMember adalah fitur yang akan diterapkan pada algoritma regresi. Setelah menentukan fitur penting yang akan digunakan, kami menguji 5 (lima) model machine learning berbeda yaitu Kneighbors Classifier, Logistic Regression, Linear SVC, Random Tree Classifier dan Random Forest Classifier. Hasil prediksi model yang diusulkan dapat dilihat pada tabel 2.

Tabel 2 Train dan Test Model Regresi

Model	Train Score	Test Score	Best Parameter
KNeighborsClassifier	84.20%	83.56%	n=3
Logistic Regression	79.09%	78.76%	Best C =4
LinearSVC	80.07%	79.92%	C=0.1
RandomTreeClassifier	82.51%	82.20%	max_depth=4
RandomForestClassifier	86.21%	84.08%	max_leaf_nodes=40

Pada tabel di atas terlihat model Random Forest Classifier menghasilkan tingkat akurasi lebih baik dari model lain, kemudian KNeighbors Classifier dengan tingkat akurasi keduanya sebesar 86% dan 84%. Selanjutnya tabel 3 menjelaskan hasil prediksi 10 pelanggan pada kedua model.

Tabel 3 Hasil Prediksi

	Person 1	Person 2	Person 3	Person 4	Person 5	Person 6	Person 7	Person 8	Person 9	Person 10
Data Aktual	Exit	Stay	Exit	Stay	Stay	Exit	Stay	Exit	Stay	Stay
kNN Classifier	Stay	Stay	Stay	Stay	Stay	Exit	Stay	Stay	Stay	Stay
Random Forest	Stay	Stay	Exit	Stay	Stay	Stay	Stay	Exit	Stay	Stay

Berdasarkan hasil prediksi ini maka dapat di simpulkan bahwa pada saat tertentu, pelanggan yang memiliki kontrak Bulan ke Bulan memiliki kemungkinan 4,85 kali lebih besar untuk melakukan churn dibandingkan seseorang yang memiliki kontrak 2 tahun, menyesuaikan dengan total biaya mereka dan pelanggan yang memiliki kontrak 2 tahun memiliki kemungkinan 5,5 kali lebih besar untuk melakukan churn dibandingkan pelanggan yang memiliki kontrak 1 tahun.

Diskusi

Temuan dari analisis data eksplorasi adalah wanita lebih cenderung churn daripada pria dan pelanggan Jerman dan yang lebih tua (berdasarkan usia) lebih cenderung melakukan churn. Proporsi yang lebih besar dari pelanggan yang berhenti memiliki nilai kredit yang buruk. Sedangkan dari lima model prediksi, random forest memprediksi pelanggan yang berpindah dengan akurasi tertinggi 86% hasil ini sangat sesuai dengan penelitian (Kaur and Kaur 2020) yang menghasilkan akurasi 85%. Dari hasil temua ini, kami memberikan rekomendasi untuk meningkatkan kesadaran di antara pelanggan tentang pentingnya menjaga skor kredit yang sehat, kemudian bagian keuangan perlu membuat penawaran investasi yang disesuaikan untuk segmen pelanggan yang berbeda dan mempertimbangkan penawaran produk dengan skema beragam untuk merayu pelanggan dari berbagai kelompok usia.

KESIMPULAN

Berdasarkan hasil pengujian pada lima (5) model machine learning untuk memprediksi churn pelanggan bank, model Random Forest Classifier dan KNeighbors Classifier lebih unggul dibandingkan model lainnya dengan akurasi masing-masing kedua model adalah 86 % dan 84 %. Sedangkan hasil analisis menemukan bahwa pelanggan wanita lebih tinggi churn dari pada pria dan tingkat umur sangat besar pengaruh terhadap churn sehingga sangat penting bagi pengambil keputusan untuk membuat segmentasi pelanggan dalam menawarkan produk untuk merayu pelanggan dari berbagai kelompok usia.

Supplementary Materials (optional)

Tidak tersedia.

Konstribusi Peneliti

Konseptualisasi, AM; Analisis formal, AM; Metodologi, AM; Administrasi proyek, MH; Pengawasan, AM; Menulis—draf asli, MH; Review & editing, AM. Semua penulis telah membaca dan menyetujui versi naskah yang diterbitkan.

Konflik kepentingan

Para penulis menyatakan tidak ada konflik kepentingan.

REFERENCES

Ahmad, Abdelrahim Kasem, Assef Jafar, and Kadan Aljoumaa. 2019. "Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform." *Journal of Big Data* 6 (1). https://doi.org/10.1186/s40537-019-0191-6.

Ahmed, Ammar A.Q., and D. Maheswari. 2019. "An Enhanced Ensemble Classifier for Telecom Churn Prediction Using Cost Based Uplift Modelling." *International Journal of Information Technology (Singapore)* 11 (2): 381–91. https://doi.org/10.1007/s41870-018-0248-3.

Almuqren, Latifah, Fatma S. Alrayes, and Alexandra I. Cristea. 2021. "An Empirical Study on Customer Churn Behaviours Prediction Using Arabic Twitter Mining Approach." *Future Internet* 13 (7): 175. https://doi.org/10.3390/fi13070175.

Ashish, Prof, Dishant Bawankule, Amit Bagde, and Falguni Deshpande. 2021. "Customer Churn Prediction in Banking Environment Using Data Science." *International Journal of Engineering and Creative Science* 4 (6): 12–17. www.ijecs.net.

Bilal Zoric, Alisa. 2016. "Predicting Customer Churn in Banking Industry Using Neural Networks." *Interdisciplinary Description of Complex Systems* 14 (2): 116–24. https://doi.org/10.7906/indecs.14.2.1.

Faris, Hossam. 2018. "A Hybrid Swarm Intelligent Neural Network Model for Customer Churn Prediction and Identifying the Influencing Factors." *Information 2018, Vol. 9, Page 288* 9 (11): 288. https://doi.org/10.3390/INFO9110288.

Hassan, Rana Saifullah, Aneeb Nawaz, Maryam Nawaz Lashari, and Fareeha Zafar. 2015. "Effect of Customer Relationship Management on Customer Satisfaction." *Procedia Economics and Finance* 23 (October 2014): 563–67. https://doi.org/10.1016/s2212-5671(15)00513-4.

Karvana, Ketut Gde Manik, Setiadi Yazid, Amril Syalim, and Petrus Mursanto. 2019. "Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry." *2019 International Workshop on Big Data and Information Security, IWBIS 2019*, 33–38. https://doi.org/10.1109/IWBIS.2019.8935884.

Kaur, Ishpreet, and Jasleen Kaur. 2020. "Customer Churn Analysis and Prediction in Banking Industry Using Machine Learning." *PDGC 2020 - 2020 6th International Conference on Parallel, Distributed and Grid Computing*, 434–37. https://doi.org/10.1109/PDGC50313.2020.9315761.

Keramati, Abbas, Hajar Ghaneei, and Seyed Mohammad Mirmohammadi. 2016. "Developing a Prediction Model for Customer Churn from Electronic Banking Services Using Data Mining." *Financial Innovation* 2 (1): 1–13. https://doi.org/10.1186/s40854-016-0029-6.

Sabbeh, Sahar F. 2018. "Machine-Learning Techniques for Customer Retention: A Comparative Study." *International Journal of Advanced Computer Science and Applications* 9 (2): 273–81. https://doi.org/10.14569/IJACSA.2018.090238.

Sayed, Hend, Manal A Abdel-Fattah, and Sherif Kholief. 2018. "Predicting Potential Banking Customer Churn Using Apache Spark ML and MLlib Packages: A Comparative Study." *IJACSA) International Journal of Advanced Computer Science and Applications* 9 (11). https://spark.apache.org/docs/latest/api/python.

Semrl, Jas, and Alexandru Matei. 2017. "Churn Prediction Model for Effective Gym Customer Retention." *Proceedings of 4th International Conference on Behavioral, Economic, and Socio-Cultural Computing, BESC 2017* 2018-Janua: 1–3. https://doi.org/10.1109/BESC.2017.8256379.

Szmydt, Marcin. 2019. *Predicting Customer Churn in Electronic Banking*. *Lecture Notes in Business Information Processing*. Vol. 339. Springer International Publishing. https://doi.org/10.1007/978-3-030-04849-5_58.

Zhao, Ming, Qingjun Zeng, Ming Chang, Qian Tong, and Jiafu Su. 2021. "A Prediction Model of Customer Churn Considering Customer Value: An Empirical Research of Telecom Industry in China." *Discrete Dynamics in Nature and Society* 2021. https://doi.org/10.1155/2021/7160527.