Laporan Akhir Tugas MK. Data Mining (KOM332), Semester Genap 2019/2020

Membangun Model Prediksi Churn dengan Data Mining

Purnama Sari (G14170008), Ragsa Endahas Ahmad (G14170025), Agus Hendra Nasution (G14170052), Bintang Rizqi Khairullah (G14170060)

Kelompok: 4, Kelas Paralel: 3

ABSTRAK

Churn merupakan permasalahan lama perusahaan jasa telekomunikasi. Perilaku pelanggan yang beralih ke perusahaan kompetitor pada industri telekomunikasi disebut *churn*. *Churn* merupakan suatu masalah bagi perusahaan jasa telekomunikasi karena mempertahankan pelanggan yang sudah ada akan lebih efisien dibandingkan dengan menarik pelanggan baru. Sehingga, tujuan dari laporan akhir ini adalah eksplorasi terhadap beberapa peubah yang terkait dengan *churn* pelanggan dan membuat sebuah model prediksi karakteristik *churn* pelanggan dengan menggunakan metode *Bagging* dengan *based learner* LGBM. Metode ini menghasilkan 10 karakteristik yang menentukan apakah seorang *customer* tersebut *churn* atau tidak dengan tingkat akurasi sebesar 72% dan F1-score kurang dari 70%.

Kata Kunci: *Bagging dengan based learner LGBM, Churn*, Perusahaan jasa telekomunikasi

PENDAHULUAN

Latar Belakang

Saat ini, alat telekomunikasi menjadi sangat penting seiring dengan berbagai perkembangan yang ada. Salah satu alasan pentingnya alat telekomunikasi ini adalah dapat merangsang pertumbuhan ekonomi secara signifikan, dan bahkan menjadi salah satu faktor keberhasilan pembangunan suatu bangsa. Tak heran, jika keberadaan telepon dan operator seluler sudah membawa masyarakat menuju kehidupan modern yang mengutamakan efisiensi dan kepraktisan. Perusahaan penyedia layanan telekomunikasi berlomba-lomba untuk memberikan pelayanan terbaik agar dipilih oleh pelanggan. Terbukanya persaingan bebas di perusahaan jasa telekomunikasi merupakan salah satu tantangan serius yang harus dihadapi oleh industri telekomunikasi. Kemudahan pelanggan untuk berpindah ke pesaing tentunya menjadi perhatian utama bagi *Customer Relationship Management* (CRM). Perilaku pelanggan yang beralih ke perusahaan kompetitor pada industri telekomunikasi disebut *churn*.

Churn pelanggan mengacu pada hilangnya pelanggan secara periodik dalam suatu organisasi. Dalam industri telekomunikasi, pelanggan menuntut produk yang lebih baik

dengan harga yang lebih murah. Sementara itu, penyedia layanan terus berfokus pada akuisisi sebagai tujuan bisnis mereka. Hal inilah yang terkadang membuat pelanggan dapat beralih operator secara periodik. *Churn* merupakan penyebab kebocoran pendapatan terbesar dari perusahaan telekomunikasi. Faktanya, merekrut pelanggan baru akan memakan biaya lima sampai enam kali lipat lebih mahal dibandingkan dengan mempertahankan pelanggan yang sudah ada. Hal inilah yang menjadi isu penting dan tantangan utama bagi perusahaan.

Oleh karena itu, manajemen *churn* menjadi senjata yang krusial dalam berkompetisi untuk melihat orientasi pilihan para pelanggan. Prediksi *churn* dapat digunakan untuk mengidentifikasi *churners* lebih awal sebelum mereka berpindah, sehingga potensi kerugian perusahaan dapat dicegah. Pemodelan *churn* dilakukan dengan menerapkan algoritma *super learner*.

Tujuan

Tujuan dari penelitian ini adalah untuk melakukan eksplorasi terhadap beberapa peubah yang terkait dengan *churn* pelanggan dan membuat sebuah model prediksi karakteristik *churn* pelanggan.

Ruang Lingkup

Data yang digunakan merupakan data sekunder yang diunduh dari laman *kaggle*. Metode yang digunakan dalam penelitian ini adalah pemodelan dengan *based learner bagging* dan LightGBM.

Manfaat

Adapun manfaat yang dapat diperoleh dari penelitian ini adalah mengetahui karakteristik pelanggan yang *churn*, sehingga dapat dijadikan bahan acuan bagi perusahaan untuk melakukan pengambilan tindakan berikutnya.

TINJAUAN PUSTAKA

Data Churn

Churning mengacu pada perilaku kehilangan pelanggan secara diam-diam dari populasi besar pelanggan. Beberapa pelanggan yang tersangkut mungkin beralih ke pesaing, sementara beberapa dari mereka mungkin meninggalkan layanan selamanya. Mungkin tidak terlihat pada awalnya, tetapi efek kerugian tersebut menumpuk dari waktu ke waktu. *Churn* pelanggan mengacu pada hilangnya pelanggan secara periodik dalam suatu organisasi (A Churi, 2015).

Bagging

Bagging merupakan merupakan salah satu metode *ensemble* dengan cara menggabungkan antara proses *bootstrapping* dan *aggregating* dalam metode pohon klasifikasi. *Bootstrapping* adalah pengambilan data contoh dari sampel yang dimiliki (resampling). *Bootstrap aggregating* atau *bagging* adalah suatu prosedur yang bertujuan untuk mengurangi keragaman yang dihasilkan dari metode statistika yang digunakan. Untuk kasus klasifikasi digunakan *majority vote* atau pemilihan suara terbanyak, sedangkan untuk masalah regresi digunakan rata-rata. Penggunaan *bagging* ini sangat membantu terutama mengatasi sifat ketidakstabilan pohon klasifikasi. Proses *bagging* dapat mengurangi galat baku dugaan yang dihasilkan oleh pohon tunggal (Hastie et al, 2008).

Extra Trees

Klasifikasi dengan metode *Extra Trees* atau yang disebut juga sebagai *Extremely Randomized Trees* merupakan varian pengembangan dari *decision tree* acak pada berbagai sub bagian dataset dan menghitung rata-ratanya untuk meningkatkan akurasi prediksi dan pengendalian *overfitting*. Berbeda dengan *Random Forest* yang pada setiap tahapannya, sampel dan keputusan diambil secara acak dan bukan diambil dari yang terbaik, *Extra Trees* membangun *group decision tree* sesuai dengan prosedur *top-down* (Geurts P, 2003). Selanjutnya, lakukan pemodelan untuk masing-masing model yang dibangun dari n data latih hasil *V-Cross Validation*. Untuk setiap model yang telah terbentuk, gunakan model tersebut untuk memprediksi data uji yang telah dibuat. Setelah itu hitung *risk* untuk setiap model. Semakin tinggi *risk* yang di dapat dari model tersebut, maka semakin kecil bobot yang diberikan untuk model tersebut. Untuk peubah respon berupa data numerik, dapat menghitung nilai *Mean Squared Error* (MSE). Jika peubah responnya berupa data kategorik, *risk* dari model dapat diketahui dengan menghitung nilai dari (1AUC). AUC di dapat dari plot sensitivitas dan 1- spesifisitas. Algoritma yang dipilih adalah dengan nilai *risk* terkecil.

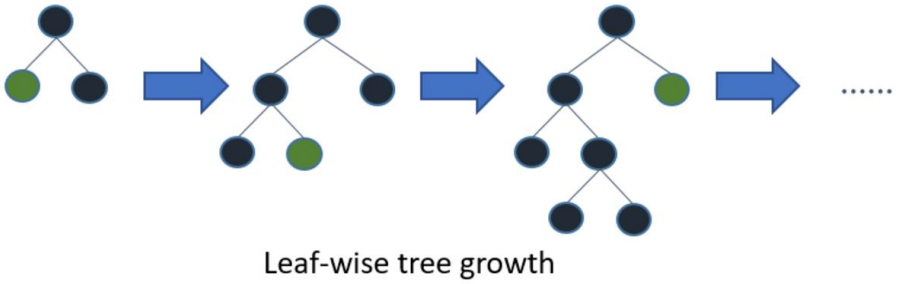
Random Forest

Random Forest (RF) adalah suatu algoritma yang digunakan pada klasifikasi data dalam jumlah yang besar. Klasifikasi random forest dilakukan melalui penggabungan pohon (tree) dengan melakukan training pada sampel data yang dimiliki. Random Forest adalah pohon dari metode *bagging* yang menggunakan pohon-pohon yang tidak berkorelasi. Sama seperti *bagging*, dibentuk banyak pohon keputusan saat proses *bootstrap* (James et al, 2003). Oleh karena itu, metode penggunaan metode Random Forest ini berupaya untuk memperbaiki proses pendugaan yang dilakukan menggunakan metode *bagging* (Breiman, 2001). Penggunaan pohon (*tree*) yang semakin banyak akan

mempengaruhi akurasi yang akan didapatkan menjadi lebih baik. Penentuan klasifikasi dengan *Random Forest* diambil berdasarkan hasil voting dari *tree* yang terbentuk.

LightGBM

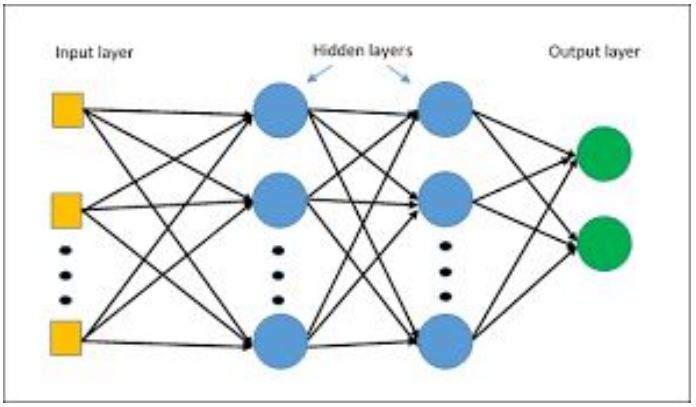
Light GBM adalah salah satu algoritma *machine learning* yang menggunakan algoritma pembelajaran berbasis pohon. Pada konsepnya LightGBM bekerja dengan membangun pohon secara vertikal, sementara algoritma lain membangun pohon secara horizontal. Artinya, LightGBM meminimalkan bias sementara algoritma lain meminimalkan variance . Ini akan memilih daun dengan max delta loss untuk tumbuh.



Gambar 1 LightGBM

Multi Layer Perceptron (MLP)

Metode klasifikasi *Multilayer Perceptron* (MLP) merupakan salah satu jenis dari algoritma jaringan syaraf tiruan yang mengadopsi cara kerja jaringan saraf pada makhluk hidup. Algoritma ini terkenal handal karena proses pembelajaran yang mampu dilakukan secara terarah. Pembelajaran algoritma ini dilakukan dengan memperbarui bobot balik (*backpropagation*). Penetapan bobot yang optimal akan menghasilkan hasil klasifikasi yang tepat. MLP terdiri dari sistem yang sederhana saling menghubungkan jaringan atau node yang diilustrasikan pada gambar dibawah. Node tersebut dihubungkan oleh bobot dan unit *output* yang merupakan fungsi penjumlahan dari *input* ke node dimodifikasi oleh transfer non-linear sederhana, atau aktivasi.



Gambar 2 *Multi Layer Perceptron*

ROC Curve (*Receiver Operating Characteristic*)

Kurva ROC adalah tampilan grafis sensitivitas (TPR) pada sumbu y dan (1 -spesivisitas) (FPR) pada sumbu x untuk berbagai titik batas nilai tes. Area di bawah kurva (*Area Under Curve* / AUC) adalah ukuran numerik untuk membedakan kinerja model, dan menunjukkan seberapa sukses dan benar peringkat model dengan memisahkan pengamatan positif dan negatif (Attenberg & Ertekin 2013). AUC digunakan untuk mengukur hasil kinerja model dan hasilnya dapat dilihat pada hasil *confusion matrix* secara manual dengan perbandingan klasifikasi menggunakan kurva ROC. Nilai maksimum AUC adalah 1, menandakan tes diagnostik sangat sempurna.

Confusion Matrix

Confusion matrix digunakan untuk evaluasi model yang dibuat. Matriks konfusi ini akan menghasilkan nilai akurasi, *precision*, dan *recall*.

- True Positive (TP) adalah jumlah pelanggan dengan status churn, ketika diprediksi hasil prediksi menunjukan churn.
- False Positive (FP) adalah jumlah pelanggan dengan status churn, ketika diprediksi hasil prediksi menunjukan non churn.
- False Negative (FN) adalah jumlah pelanggan dengan status non churn, ketika diprediksi hasil prediksi menunjukan churn.
- True Negative (FN) adalah jumlah pelanggan dengan status non churn, ketika diprediksi hasil prediksi menunjukan non churn.

Recall dihitung untuk mengevaluasi seberapa besar *coverage* suatu model dalam memprediksi suatu kelas tertentu. *Recall* didapatkan dengan menghitung perbandingan antara jumlah data untuk satu kelas tertentu yang diprediksi dengan benar dibagi jumlah total kelas tersebut. *Precision* dihitung untuk mengevaluasi seberapa baik ketepatan model dapat memprediksi suatu kelas. Ini didapatkan dengan menghitung perbandingan antara jumlah data untuk satu kelas tertentu yang diprediksi benar, dibagi dengan jumlah total prediksi kelas tersebut. Selain nilai *recall* dan *precision,* dapat juga dihitung nilai *F1-measure* yang merupakan kombinasi antara nilai *recall* dan *precision*. Inilah yang biasanya dijadikan sebagai akurasi pada pengujian.

METODE

Data

Data yang digunakan adalah data sekunder yang berasal dari situs web *kaggle.* Data ini terbagi menjadi 2, yaitu data latih dan data uji dengan masing-masing berformat *csv*. Data churn ini terdiri dari 58 atribut. Data latih terdiri dari 51.047 observasi sedangkan

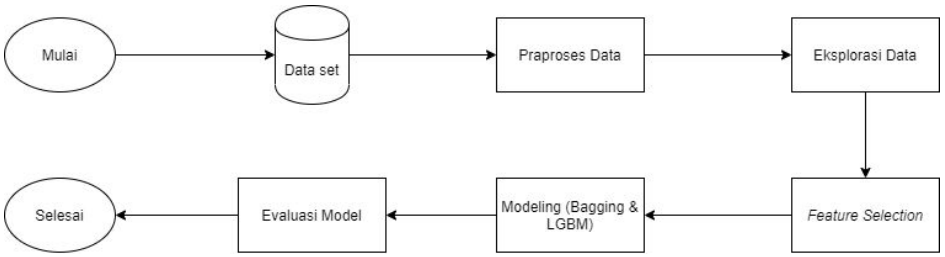
data ujinya sebanyak 20.000 observasi. Berikut adalah beberapa atribut yang terdapat pada data churn yang digunakan.

Tabel 1. Beberapa Atribut / Peubah Sebelum Tahap Pra- proses

No	Nama Atribut	Tipe	Penjelasan
1	CurrentEquipmentDays	float64	Pelayanan yang dibutuhkan customer tiap hari
2	PercChangeMinutes	float64	Persentase lama penggunaan layanan telekomunikasi
3	MonthlyMinutes	float64	Jumlah menit panggilan yang dihabiskan per bulan
4	PercChangeRevenues	float64	Pendapatan dari penggunaan setiap menit
5	MonthlyRevenue	float64	Pendapatan bulanan customer
6	PeakCallsInOut	float64	Panggilan keluar dan masuk
7	OffPeakCallsInOut	float64	Panggilan keluar dan masuk tidak tersambung
8	UnansweredCalls	float64	Panggilan tidak terjawab
9	ReceivedCalls	float64	Panggilan diterima
10	OutboundCalls	float64	Panggilan keluar

Tahapan Kegiatan

Penelitian ini dilakukan dalam beberapa tahapan, yaitu *data selection* , *data pre-processing* , *data transformation*, eksplorasi data, *feature selection* dan kemudian dilakukan modeling menggunakan algoritma *bagging* dan LightGBM. Setelah itu, dilakukan interpretasi hasil dan evaluasi terhadap model yang diperoleh. Secara sistematis, alur dari tahapan kegiatan dapat dilihat pada gambar 2.



Gambar 3 *Flow chart* tahapan kegiatan

Lingkungan Pengembangan

Perangkat keras yang digunakan dalam penelitian ini adalah komputer personal dengan spesifikasi sebagai berikut.

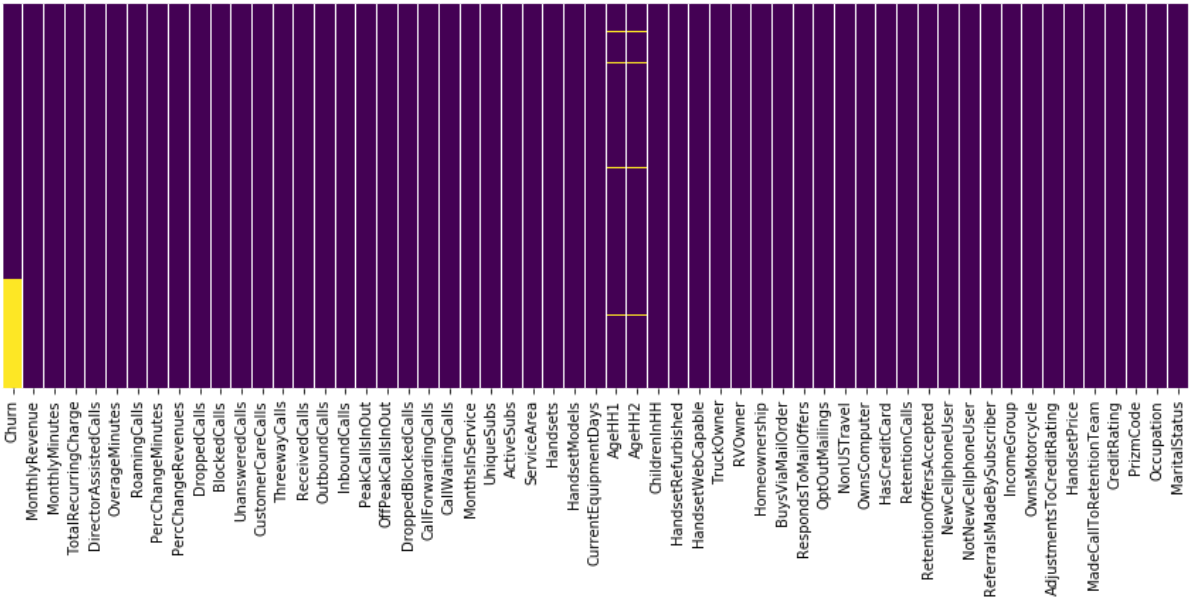
a. Prosesor : Processor intel core i5
b. Memory : 4GB
c. VGA : NVIDIA GEFORCE 940M 2GB

Adapun perangkat lunak yang digunakan dalam penelitian ini adalah bahasa pemrograman Python

HASIL DAN PEMBAHASAN

Pra Proses Data dan Transformasi Data

Tahapan awal dalam penelitian mengenai Model Prediksi *Churn* ini adalah praproses data atau proses mempersiapkan data agar dapat dianalisis. Penelitian ini menggunakan dataset "*Telecom Churn*" yang terdiri dari 58 atribut yang terbagi ke dalam data latih dan data uji. Pada masing-masing data, terdapat beberapa data yang hilang (*missing value*) yang harus diisi dengan suatu nilai agar analisis data dapat dilakukan.

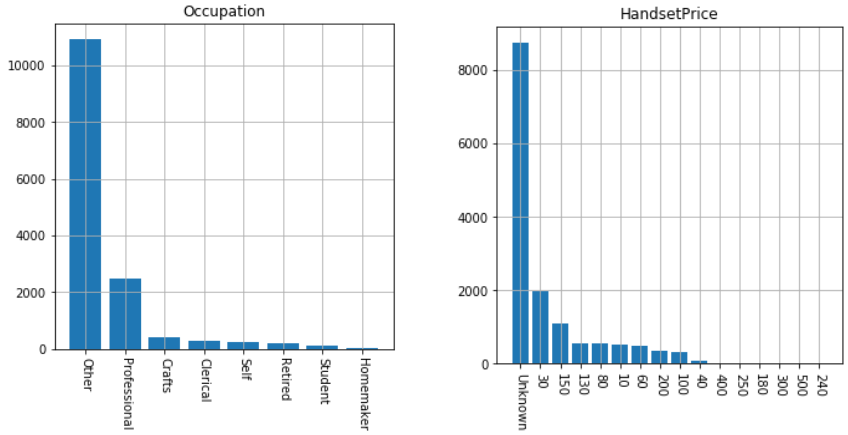


Gambar 4 Visualisasi *missing value*

Pengisian *missing value* dilakukan berdasarkan skala pengukuran dari atributnya. Pada kasus ini, pengisian data hilang dilakukan dengan nilai median data. Selain pengisian *missing value*, dari 58 atribut tersebut terdapat beberapa peubah dengan tipe data *object*, sehingga perlu dilakukan transformasi agar proses analisis lanjutan dapat dilakukan. Perubahan tipe atribut ini dilakukan dengan *label encoder* dari 23 tipe data object diubah menjadi numerik, sehingga total atribut sebanyak 850 atribut.

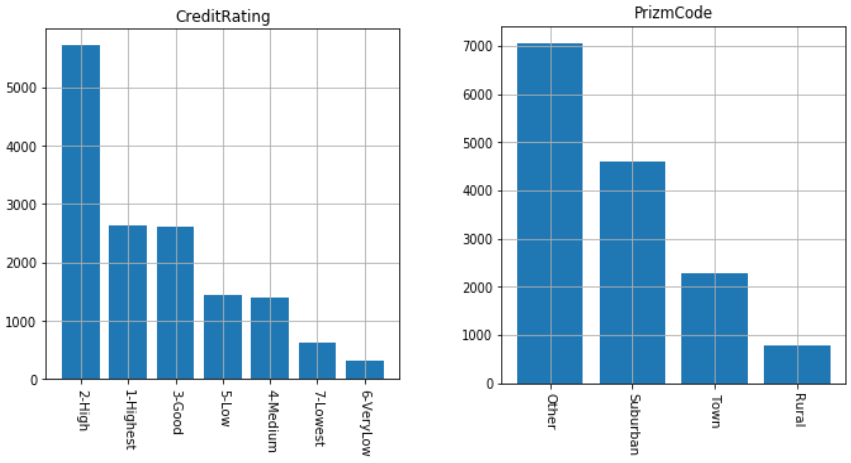
Eksplorasi Data

Eksplorasi data dilakukan untuk mendapatkan informasi mengenai perilaku konsumen yang churn, sehingga hanya dipilih beberapa atribut yang dianggap lebih menggambarkan karakteristik pelanggan.



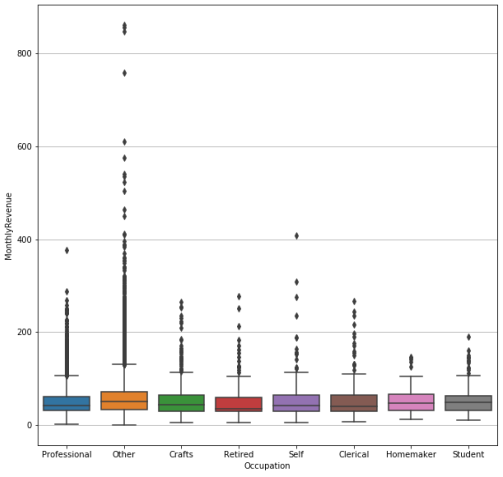
Gambar 5 (a) Visualisasi karakteristik pelanggan *churn* peubah Occupation dan (b) Visualisasi karakteristik pelanggan *churn* peubah HandsetPrice

Plot diatas menunjukan karakteristik konsumen yang memiliki latar belakang pekerjaan (*occupation*) dengan "other" cenderung melakukan *churn*. *Customer* yang paling jarang melakukan *churn* adalah pelanggan dengan profesi ibu rumah tangga. Artinya, ibu rumah tangga memang biasanya hanya menggunakan alat telekomunikasi untuk hal-hal biasa seperti telepon, sehingga kegiatan berpindah provider akan sangat jarang dilakukan. Disisi lain konsumen yang memiliki HandsetPrice kurang dari 100 lebih banyak melakukan *churn*. Hal ini bisa terjadi karena orang - orang profesional akan cenderung memilih provider yang lancar sesuai kebutuhan bisnisnya dan untuk konsumen dengan HandsetPrice rendah menggambarkan tingkat pendapatan yang dimiliki, bisa jadi tingkat pendapatan lebih rendah dari pada pengeluaran sehingga lebih cenderung melakukan *churn*.



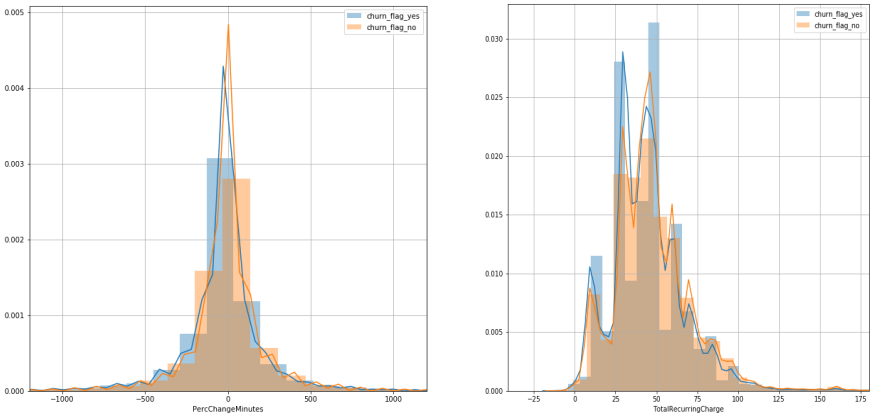
Gambar 6 (a) Visualisasi karakteristik pelanggan *churn* peubah CredRating dan (b) Visualisasi karakteristik pelanggan *churn* peubah PrizmCode

Berdasarkan plot diatas dapat dilihat bahwa pelanggan yang memiliki CreditRating yang tinggi cenderung melakukan *churn*. CreditRating menggambarkan besar peluang konsumen untuk membayar hutangnya tepat waktu. Semakin tinggi nilai CreditRating konsumen, maka diasumsikan bahwa konsumen tersebut memiliki pendapatan yang tetap, dapat diasumsikan bahwa semakin tetap penghasilan konsumen maka semakin tinggi kemungkinannya untuk melakukan *churn*. Hal lain yang dapat disimpulkan adalah bahwa pelanggan yang memiliki PrizmCode pada daerah *suburban* memiliki kecenderungan untuk melakukan *churn* dibanding daerah *town* dan *rural* mungkin hal tersebut dapat menggambarkan hubungan tempat tinggal konsumen dengan kecenderungan konsumen untuk melakukan *churn* , diduga penduduk yang tinggal didaerah *suburban* lebih cenderung untuk melakukan *churn.* Artinya, pelanggan yang tinggal di daerah pinggiran kota atau tidak jauh dari pusat kota lebih cenderung untuk melakukan *churn*, jika dibandingkan dengan pelanggan yang bertempat tinggal di daerah pedesaan yang jauh dari berbagai akses dan fasilitas kota.



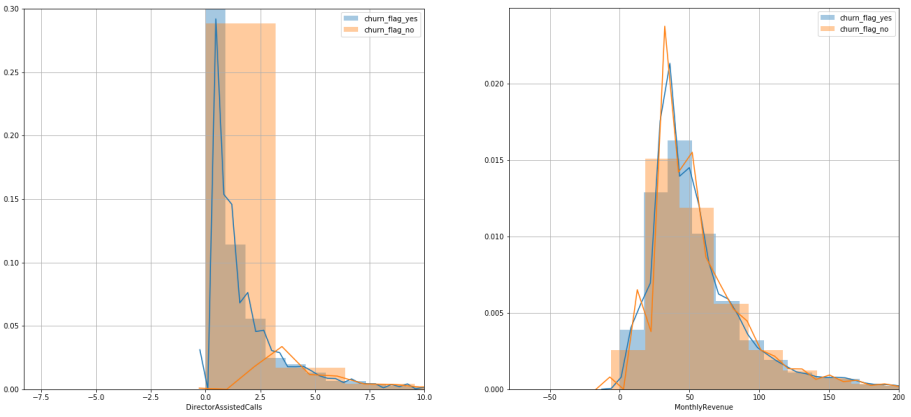
Gambar 7 Box plot atribut occupation vs MonthlyRevenue

Dilihat dari Gambar 7 dapat dikatakan bahwa *customer* dengan berbagai jenis profesi memiliki jumlah pendapatan bulanan yang beragam, dan tentunya terdapat banyak pencilan. Dari beberapa profesi tersebut, terlihat bahwa profesi dengan rata-rata pendapat bulanan paling kecil dibandingkan profesi lain adalah para pensiunan (retired) diikuti oleh profesi tokoh agama (clerical).



Gambar 8 (a) Visualisasi sebaran data *churn* peubah PercChangeMinutes dan (b) Visualisasi sebaran data *churn* peubah TotalRecurringCharge

Dilihat dari Gambar 8 (a) dan (b), data pada peubah PercChangeMinutes terlihat menyebar normal, namun antara kategori pelanggan *churn* dan tidak pada plot tersebut tampak memiliki perbedaan bentuk pola data. Artinya, bisa dikatakan bahwa peubah PercChangeMinutes ini cukup berpengaruh dalam mengkategorikan pelanggan *churn* atau tidak. Untuk data pada peubah TotalRecurringCharge, tampak bahwa data cenderung menjulur ke kanan. Namun, antara kategori *churn* dan tidak *churn* pada atribut ini memiliki pola data yang hampir mirip sehingga dapat dikatakan bahwa peubah TotalRecurringCharge tidak berpengaruh signifikan dalam membedakan *customer churn* atau tidak. TotalRecurringCharge menunjukkan seberapa banyak total biaya yang dihabiskan *customer* untuk biaya pembelian ulang terhadap suatu provider telekomunikasi.



Gambar 9 (a) Visualisasi sebaran data *churn* peubah DirectorAssistedCalls dan (b) Visualisasi sebaran data *churn* peubah MonthlyRevenue

Berdasarkan plot sebaran sebaran pada Gambar 9 (a) diatas dapat disimpulkan bahwa pemakaian layananan DirectorAsisstedCalls tidak terlalu berpengaruh dalam pengkategorian pelanggan *churn* atau tidak karena antara keduanya memiliki pola sebaran data yang hampir mirip. Sedangkan untuk atribut MonthlyRevenue memiliki pola sebaran data yang cenderung menjulur ke kanan yang artinya, hanya sedikit pelanggan dengan pendapatan bulanan tinggi. Berdasarkan Gambar 9 (b), terlihat bahwa antara kategori *churn* dengan tidak *churn* memiliki pola sebaran data yang cenderung berbeda. Artinya, atribut MonthlyRevenue ini berpengaruh signifikan dalam pengkategorian pelanggan *churn* atau tidak.

Modeling

Proses klasifikasi "Customer Churn" dicoba dengan menggunakan beberapa metode, seperti Extra Trees, Random Forest, Voting Classifier, MLP, Bagging dan LightGBM yang dikombinasikan dengan boosting agar meningkatkan nilai akurasi untuk semua data. Tentu saja data yang akan diklasifikasikan harus sudah dirapikan terlebih dahulu seperti, mengubah peubah kategorik menjadi peubah boneka serta memilih peubah-peubah yang akan dimasukkan ke dalam model sesuai dengan hasil dari *feature selection*. Berikut hasil prediksi dari masing - masing model :

Tabel 2. *Result set* Algoritma

No	Model	Akurasi	Presisi	Recall	F1 Score
1	LGBM	0.7088	0.5117	0.1981	0.5114
2	Voting Classifier	0.7139	0.5555	0.1333	0.5200
3	MLP	0.7079	0.5155	0.1028	0.4971
4	Bagging + LGBM	0.7224	0.5661	0.1363	0.5254
5	AdaBoost + RF	0.7180	0.5958	0.0717	0.4798
6	AdaBoost + Extra trees	0.7180	0.4913	0.0511	0.4602

Nilai akurasi, presisi, recall, *F1-score* pada model di atas ini cukup besar dan hasil dari *accuracy* diatas 70% untuk semua model yang artinya nilai data yang diujikan cukup baik untuk memprediksi data aktual yang cukup bervariasi, tetapi tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi juga kurang baik, dapat dilihat dari nilai *recalls* dibawah 70%. Baik nilai akurasi yang tinggi tetapi nilai *recall*, presisi, dan *F1-score* yang rendah mengindikasikan data yang dimiliki masih kurang bagus atau teknik

yang kami gunakan kurang tepat. Mengingat nilai akurasi sebesar 72% untuk *bagging model* sudah cukup bagus pada kasus ini.

Evaluasi Model

Untuk mengevaluasi model akan dilakukan perbandingan nilai *Cross Validation* (CV) dari masing - masing model. Berikut adalah hasil akurasi dari CV dengan k-fold 5.

Tabel 3. Hasil *cross validation* algoritma

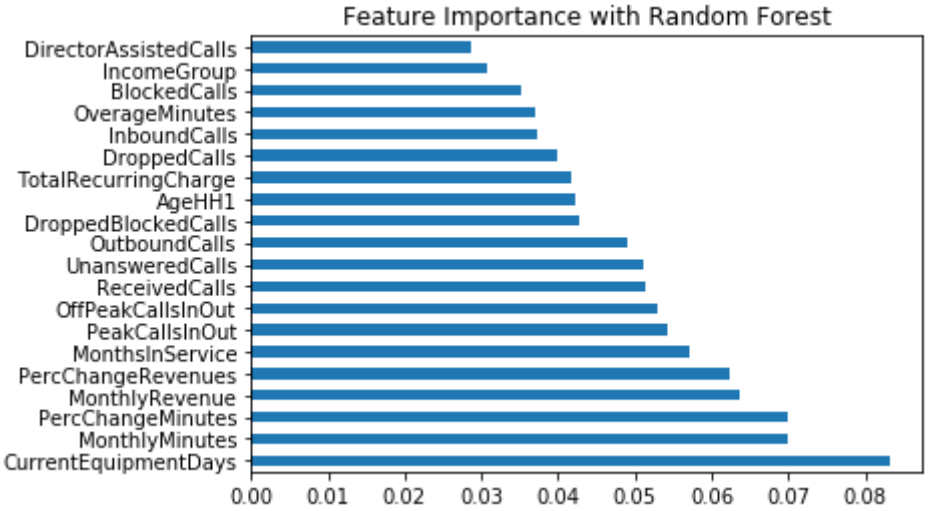
No	Model	CV 1-Fold	CV 2-Fold	CV 3-Fold	CV 4-Fold	CV 5-Fold	CV Mean
1	LGBM	0.70728977	0.70728977	0.71396586	0.70878813	0.71060733	0.7095
2	Voting Classifier	0.71974255	0.71372604	0.72124265	0.71592499	0.71382592	0.7168
3	MLP	0.71064782	0.70435148	0.70766863	0.70822838	0.68205989	0.7025
4	Bagging + LGBM	0.72100182	0.71540507	0.72390148	0.71914358	0.72082284	0.7200
5	AdaBoost + RF	0.71456555	0.7186232	0.71578505	0.71298629	0.71550518	0.7154
6	AdaBoost + Extra trees	0.7112075	0.71596474	0.71410579	0.71116709	0.7152253	0.7135

Berdasarkan hasil Cross Validation diatas didapatkan bahwa tingkat *accuracy* model *bagging* adalah yang tertinggi dengan nilai rata - rata CV sebesar 0.72. Model *cross validation* ini berguna untuk mengatasi masalah overfit. Jika dilihat nilai akurasi model dengan CV tidak berbeda jauh berada pada kisaran 72%.

Pemilihan Model Berpengaruh

Berdasarkan hasil pra-proses yang dilakukan pada peubah data *customer churn* diperoleh 58 peubah bebas yang akan digunakan untuk melakukan klasifikasi. Melalui algoritma Random Forest dengan nilai CV dan *accuracy* tertinggi, kita memperoleh 20

peubah dari 58 peubah yang dianggap berpengaruh terhadap peubah target. Peubah tersebut ditampilkan pada gambar dibawah.



Gambar 9 *Output Feature Importance*

Untuk modeling kami hanya mengambil 10 peubah yang memiliki nilai proporsi lebih besar dari 0.05. Peubah tersebut adalah 'CurrentEquipmentDays', 'PercChangeMinutes', ‘MonthlyMinutes', 'PercChangeRevenues', 'MonthlyRevenue', ‘'PeakCallsInOut', 'OffPeakCallsInOut', 'UnansweredCalls', 'ReceivedCalls', dan 'OutboundCalls'.

KESIMPULAN DAN SARAN

Hasil analisis mengenai pola perilaku konsumen didapatkan 10 karakteristik yang menentukan apakah seorang *customer* tersebut *churn* atau tidak. 10 atribut yang dimiliki digunakan untuk melakukan pemodelan menggunakan model *Bagging* dengan *Based learner* LGBM, didapatkan tingkat akurasi sebesar 72% yang artinya model sudah cukup baik dalam membedakan *customer churn* atau tidak, tetapi melihat nilai recall, precision, dan F1-score kurang dari 70% mengindikasikan model masih mengalami bias atau overfit. Hal ini bisa diatasi dengan menggunakan metode *factor analysis* atau menggunakan *resampling* data sehingga pengaruh bias bisa diminimalisir.

DAFTAR PUSTAKA

Attenberg, J., & Ertekin, S. 2013. Class Imbalance and Active Learning. In H. He, & Y. Ma, Imbalanced Learning: Foundations, Algorithms, and Applications (pp. 101-149). New Jersey: John Wiley & Sons. [diunduh 2020 Jun 1].

Breiman, L. 2001. Random Forests. *Machine learning,* 45(1), pp.5-32. University of California Berkeley.

Churi, A., Divekar, M., Dashpute, S., & Kamble, P. 2015. Analysis of Customer Churn in Mobile Industry using Data Mining. International Journal of Emerging Technology and Advanced Engineering, 5(3): 225-230. Retrieved from www.ijetae.com/files/Volume5Issue3/IJETAE_0315_41.pdf [diunduh 2020 Jun 1].

Geurts, P. 2003. Extremely randomized trees. Technical report. University of Liege: Department of Electrical Engineering and Computer Science. https://www.researchgate.net/publication/220343368_Extremely_Randomized_Trees [diunduh 2020 Jun 2].

Hastie, T. J., Tibshirani, R. J. and Friedman JH. 2008. The Elements of Statistical Learning: Data-Mining, Inference and Prediction. 2nd ed. New York: Springer Verlag.

James, G., Witten, D., Hastie, T., & Tibshirani, R. 2013. An Introduction to Statistical Learning. New York : Springer Verlag.