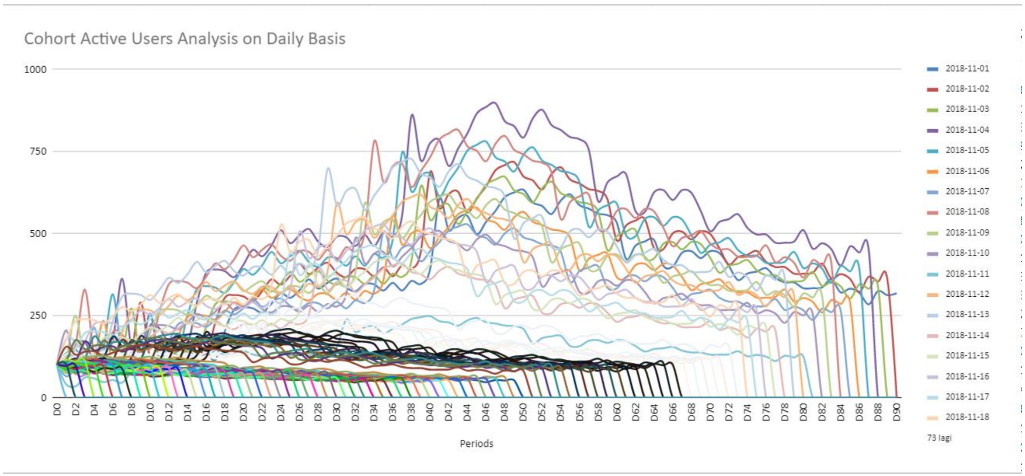
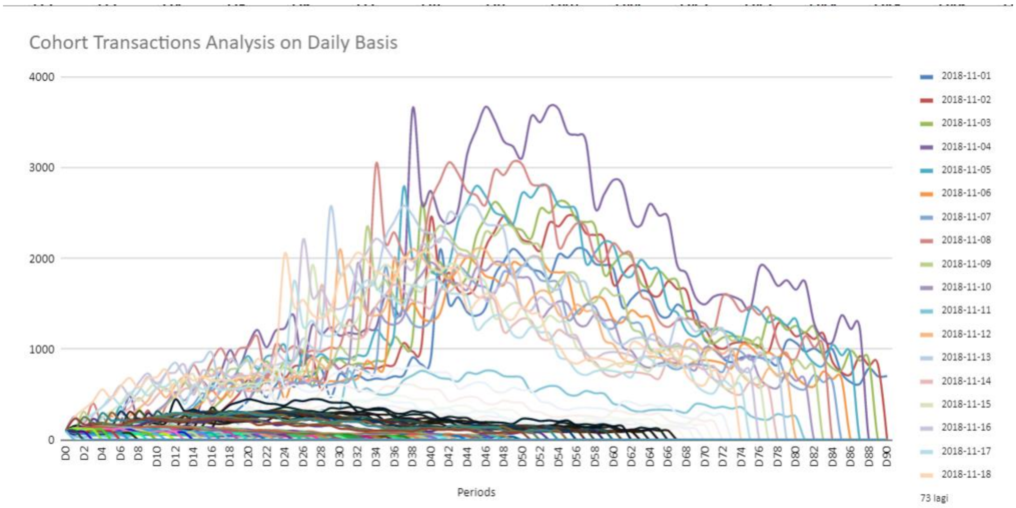
Report of Dana's Data Science Assignments

The problem on this assignment is a company wants minimizing reviving customer which are already churn or in dormant segment because it will be spending more cost than new customer, so for preventing on that situation, company wants to predict a customer who are going to churn specifically on 1st march 2019. In the assignment, flag churn is 20 days if a customer didn't do transaction for more than 20 days since last transactions. I wonder, does 20 days flag churn is already suitable for a company or we need to redefine number of days to flag as a churn, in this analysis, I already provided the analysis with retention cohort analysis by using cohort transactions analysis and cohort active user analysis. The following graph of cohort analysis be provided below.





Based on cohort graph above, we can gather insight is that the retention transactions and active users have increasing trends for 40 to 50 days after their last transactions, it started to decreasing from days 51. It means that 20 days is short to flag a customer churn as we can see on the graph the retention transactions and active user will be dropped stars from days 51, I think based on graph above, we could redefine flag churn if a customers didn't do transaction for more than 51 days.

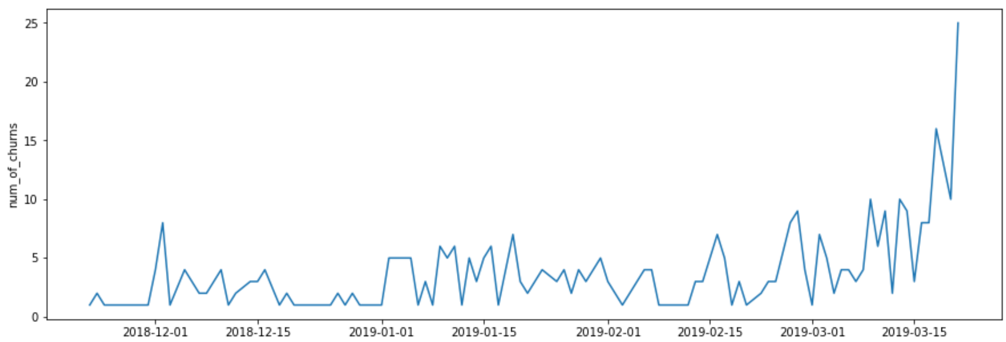
Moving on to the request of the assignment is predict a customer who are going to churn specifically on 1st march 2019. There were several steps to predict customer churn by building models. First, aggregate transactions data into user level to create new features or we called it feature engineering (already provided on the notebook). Second, Create flag churn in training and testing dataset, training dataset flag churns based on last_created_time on order_data minus last_date on January means 2019-01-31, while testing dataset using last_trx_data on test_data minus February 28, then we got a flag churn both dataset (training and testing). Third, do some feature selection by applying optimum bining, to find which features can distinctive population of churn and non-churners, here is the result analysis for every features as follow:

feature	IV	is monotonic?	Level	Remove?
user_id				
monthly_sum_mrcht	0,49	yes	strong predictor	no
daily_avg_trx	0,61	yes	suspicious	no
monthly_avg_trx	1,57	yes	suspicious	no
daily_avg_ord_amt	0,16	no	medium predictor	yes
monthly_avg_ord_amt	1,31	yes	suspicious	no
daily_avg_prm_amt	0,25	yes	medium predictor	no
daily_avg_net_amt	0,13	no	medium predictor	yes
monthly_avg_net_amt	1,31	yes	suspicious	no
nopromo_trx	2,26	yes	suspicious	no
coupon_trx	0	no	useless	yes
spinwheel_trx	0	no	useless	yes
diff_avg_seconds	0,52	yes	suspicious	no
is_premium_user	0,03	yes	weak predictor	no

The Blue colors on left table above denoted features that should be removed when build a model because based on optimum bining, it is no monotonic (decreasing or increasing) and Information Value is small. Forth, Build some modeling for choosing which one is the best, here is the summary model as follow:

	model	NLL_Train_score	NLL_Test_score	AUC_Train_score	AUC_Test_score	AUC_delta_score	times	weighted_score
5	sgd	11.888454	12.33316	0.67486	0.669118	0.005741	23.071675	0.132461
0	Decision Tree	0.309272	1.263708	0.927847	0.810888	0.116959	2.340402	0.273731
1	randomforest	0.150174	0.920697	0.995504	0.817529	0.177975	3.852738	0.68569
4	logreg	0.554588	0.794116	0.787063	0.719975	0.067087	13.606713	0.912839
6	svm	0.551906	0.540781	0.850305	0.805862	0.044443	25.70354	1.142343
2	bagging	0.229416	0.827262	0.970119	0.819809	0.150309	10.809502	1.624769
3	adaboost	0.005611	8.065375	0.999959	0.802955	0.197003	11.152341	2.19705
7	catboost	0.236489	0.668337	0.962858	0.820677	0.14218	25.989705	3.695227

I select svm model as final model because it has consistency score in both evaluation score (NLL and ROC). Fifth, Predict date flag churn for each customer, here is the graph of date churn for each customer as follow:



For overall predicted churn on testing data around 98 customers which divided for each date, but for specifically churn on 1st march, it will be only 1 customer who are going to churn.

Lastly, it is additional analysis to which segment that should be retain or not by applying risk rank table as follow:

Grade	Proba MIN	Proba MAX	NoA Population Total	NoA Churns Total	NoA Population Rate (%)	NoA Churns Rate (%)	NoA Churns Rate Cumulative (%)	Amt Population Total	Amt Population Cumulative	Amt Churns Total	Amt Populations Rate (%)	Amt Churns Rate (%)	Amt Churns Rate Cumulative(%)
A1	0,240	0,250	117	2	10%	2,0%	2,0%	88184188	88184188	1051280	18%	1,00%	1,0%
A2	0,250	0,260	116	6	20%	5,0%	3,0%	114730778	202914966	4990260	41%	4,00%	3,0%
A3	0,260	0,270	117	8	30%	7,0%	5,0%	54050069	256965035	3981761	52%	7,00%	4,0%
A4	0,270	0,280	116	19	40%	16,0%	8,0%	41280187	298245221	5507306	60%	13,00%	5,0%
B1	0,280	0,300	117	22	50%	19,0%	10,0%	101303690	399548911	6891913	81%	7,00%	6,0%
B2	0,300	0,330	116	36	60%	31,0%	13,0%	27243894	426792805	11305814	86%	41,00%	8,0%
B3	0,330	0,380	116	46	70%	40,0%	17,0%	22872501	449665305	9298055	91%	41,00%	10,0%
C1	0,380	0,430	117	74	80%	63,0%	23,0%	33758657	483423962	31067981	98%	92,00%	15,0%
C2	0,430	0,530	116	56	90%	48,0%	26,0%	7039399	490463361	3274917	99%	47,00%	16,0%
C3	0,530	0,670	117	82	100%	70,0%	30,0%	4358445	494821806	3005717	100%	69,00%	16,0%

Based on table above, The green one, we won't offer them with retained product, but we can give them a product for upgrade level. The grey one, we are not going to offer this segment as we need to customized product for each customer on this segment to make customer retain, but it will be spending more money. I select the yellow ones to get offered of retain product because churns rate for each segment is still less than 50%.