

note

114 views

Thoughts on Customer Selection

I'm intrigued by the nuance that we are looking for customers that will never pay their bill not ones that might be delinquent and may still be delinquent next month but will eventually pay.

This suggested to me a classification model that looks at one set of data to predict a longer term outcome.

My think out loud approach is a supervised learning approach where we look at historical data through time period x and then classify people who paid their bill or began paying in $x + t$ vs people who remain unpaid in $x + t$. (i.e. $t = 6$ months)

So for example look at your target variables on 12/31/17 for all customers. All customers who are current on their bill and have never been late can be eliminated. This has the advantage of substantially reducing the data to be looked at.

Now I would assign a binary categorical value to the customer of 1 if they remained delinquent in $x+t$ (six months) and 0 if they paid all or a portion of their bill (at some reasonable cutoff) indicating they are likely to continue to try to pay their balance thus trying to capture those variables that are predictive of sustained delinquency. Perhaps some analytics on the optimal value for t . The result of our model is applied to the current population with their data through today.


The variables on 12/31/17 for learning I would initially like to include are length of delinquency, age of account, amount of delinquency, rent/own, fico, age, home value. Obviously there would be some analysis about which variables are best predictors so I would use LASSO for variable selection and I like an SVM model because of the ability to adjust the C value to weigh the model towards selecting very likely delinquents while accepting false positives (keep power on) at more marginal cases.

hw8

Updated 1 day ago by Alexa Langford and Gregory Trautman


followup discussions for lingering questions and comments

☒ Resolved ☐ Unresolved

 **Sam Marquez** 1 day ago

I'm a bit confused on the first portion - where you're talking about the time period. This would seem to suggest a time-series model, but I don't think that's what you're suggesting because it wouldn't make much sense. At least at the point where we want to know pay/not pay. Are you suggesting using the $x+t$ period to classify whether they will pay or not pay? If so, I don't think we'd need to do that, given the model will determine that coefficient for you. i.e. if you run a regression and one of your variables is $x_{\text{timeunpaid}}$, the regression will give you a coefficient for $x_{\text{timeunpaid}}$ and how strongly that correlates to the pay/not pay target variable. Basically, you don't need to worry about defining the time period as it's handled for you.

☒ Resolved ☐ Unresolved

 **Kelly Lester** 1 day ago

I was thinking the same approach as I believe Alexa posted. Supervised learning using (Logistic regression) to determine the significant factors to why people don't pay. Variables such as recent credit score changes, age, is the property rented or owned. Also, external trends such as weather, significant economic changes such as recession, unemployment, time of year (very cold or very hot trends) may be helpful. The external factors would likely be added after the supervised learning of customer data.

Thanks



Gregory Trautman 1 day ago Thanks for the follow up What I was trying to capture was a chronic non payer but I agree with your response that a variable for total months unpaid can accomplish the goal and may be more efficient.

☒ Resolved ☐ Unresolved

 **Jordan Nelson** 1 day ago

Thinking about all of the factors (particularly those related to demographics/income/credit) required to model customer paying/nonpaying behavior makes me think it would be very costly, time-consuming, and potentially tricky to make the case for why the power company should have access to that data (especially in the age of GDPR).

This makes me wonder if it might be possible to model customer paying/nonpaying behavior with a probability distribution. Perhaps we could use try using the geometric distribution to help model how many months/bills/cycles people who haven't paid their most recent bill but will in the future typically go before remembering and paying again -- or the Weibull distribution to help model this as the time between missing a payment and paying again?



Sandip 16 hours ago Are you thinking Weibull for a given customer or for the whole customer base? I wonder if there will be enough data to make a probability distribution for a single customer.



Jordan Nelson 15 hours ago Thanks for adding your thoughts! Yeah, maybe I'm way off here, but I wonder if you could take historical data for customers who have missed payments and later resumed paying and test if it fits geometric (for something like cycles until a lapsed customer resumes payment) or Weibull (for something like amount of time until a lapsed customer resumes payment). And if so, that should make it easier to decide when it's safe to predict that a lapsed customer likely won't resume payment. So pretty similar to the lecture video example of trying to use probability distributions to determine when it's safe to say a season ticket holder likely won't show up to a game so you can sell their seat as an upgrade.



Gregory Schreiter 8 hours ago I was thinking something similar. I would like to include some sort of simulation aspect to this project, and perhaps we could model the 3 categories of delinquent payers by some sort of weibull/geometric distribution based on whether we model continuous time or number of payments until delinquency. Then we could combine this information (prioritizing those who are able to pay but won't) in a simulation model containing geospatial data to determine how many workers/trucks we would need throughout the city to do this task. Got to start writing now... this course is relentless! o_o