RANKMINER PROJECT QUERIES Week 1

MORNING BATCH

Note that the numbering below doesn’t correspond exactly to the numbers in the original document; however, the questions were answered in order so finding an answer to a particular question can be done by a rudimentary search. For the Evening Batch, I tried to refer to the appropriate answer number in the Morning Batch if the question was a repeat; if the question was a repeat from another one in the Evening Batch however, it was just skipped as was done with the Morning Batch.

1. Explanation of rec\_statuses:
   1. PTP arranged – “promise to pay” indicates the debtor’s intention to provide payment on debt; considered a positive outcome.
   2. CUST 3 – call transfer (this only implies a Right Party Contact)
   3. CUST 4 – transfer to closer indicates the debtor has promised to make a payment over the phone and is being transferred to an agent who can process the payment
   4. CUST RPC PTP 1 – alias for promise to pay
   5. Customer hung up – a physical hang-up on the agent, most likely before any other resolution could be reached (this only implies a Right Party Contact, but if it is, it is considered a negative outcome).
   6. Debtor dispute – when approached about a debt, debtor denies the charge, requiring the collection agency to provide proof of debt, usually through sending paperwork in the mail. This is considered an “unresolved” call.
   7. DNC – Do Not Call indicates a debtor’s complete refusal to cooperate with the collection agency. This is a “refusal to pay”, or a negative outcome.
   8. Operator transfer – same as CUST 3.
2. PAYROLL\_ID is a unique identifier for an agent. Note that agent\_id is also a unique indicator; payroll\_id was included in the data for verification between original data sources and is not strictly needed. “src1” and “src2” mean “source 1” and “source 2”, again indicating the data source for those entries (both sources are from HR). Payroll\_ids across the 2 sources should be the same; if they are not, it represents an error in the data.
3. TERM\_DATE indicates the date on which an agent was terminated (either was fired or resigned). As with payroll\_id, “src1” and “src2” indicate the original data source. Both are included to establish the veracity between sources. If an agent has different dates across the 2 sources, it indicates an error in the data. The true date might be ascertained by looking at the dates of calls for that agent in the call data.
4. WORK\_SHIFT indicates what time of day the agent worked. “AM”/”am” and “PM”/”pm” are regular shifts, “MOD” shifts have the agent working 1 day over the weekend in lieu of 1 weekday, “PT” is part-time, “SS” indicates a split shift, which is full-time but working within multiple groups over the course of the month, and “WE” is the weekend shift, 10 hours/day, Friday through Monday. Work shifts are provided across the 2 different sources for the sake of veracity. Differences between the 2 for an agent indicate an error in the data & verification of the agent’s true working hours might be ascertained from the call data.
5. Regarding agent wages, agents are paid on an hourly basis and the wage is indicated by the <month>\_hourly\_rate which is in $/hr. The agent’s shift indicates the total number of hours s/he must work for any given month. The hours for a shift vary from month to month because of the number of days in the month and any holidays in that month; regardless, AM & PM (including MOD) shifts & the split shift tend to be within 165-185 hours, WE and PT shifts tend to be roughly 20-40 hours less. Regardless of a month’s number of hours for a particular shift, you can assume that they are all in the same proportion to each other across all months. The total number of hours to be worked for a month can usually be derived from the agent data by finding the mode of the <month>\_hrs\_worked among agents. The <month>\_hrs\_worked indicates the actual hours an agent worked (i.e. doesn’t include PTO, sick leave, jury duty, etc.). The agents work directly for the collection agency (recall that the collection agency purchases debt from companies and that agents get assigned to a certain company’s debt, but they are paid by the agency).
6. GROUP indicates which client’s debt the agent is assigned to take calls for. “Auto” are for vehicle loans; “CC” is credit card debt; “COMM” is for “commercial” debt; “TMOBILE” is a cell carrier; “Legal” are outstanding legal fees; “CIN BELL” is Cincinnati Bell, a phone company; “ATT” is AT&T, a cell carrier; “FM” is unspecified in the available data; “Telecom”/“TELECOM”/”telecom” is referred to as “Debt Buy Telecom”, most likely any other telecom/cell provider not explicitly listed; “DISH” is DISH network, a broadband TV provider; “DIRECTV” is also a satellite-based broadband TV provider; “Verizon” and “Sprint” are cell carriers; “INB” refers to inbound calls which typically cannot be categorized by client, with the exception of “INB DSH” and “INB SPR” which are inbound calls for DISH network and Sprint, respectively.
7. AGENT\_ID is the primary unique indicator for a particular phone agent – it’s the one that should be used to identify an agent across data sources. The three letters are derived from the agent’s name.
8. TERM\_CODE is the piece of data in the original data sources which indicated the type of termination. For convenience, the columns TERM\_TYPE and TERM\_REASON are provided which indicate the code’s meaning. The first letter of the code indicates whether the agent chose to leave or not (voluntary or involuntary, respectively), and the second letter indicates the reason the agent left. The description for termination reason was given as what is indicated in the agent data. Most of the values should be self-explanatory. From my own experience, “insubordination” indicates a failure to adhere to the collection agency’s employment policies, “behavior” can indicate poor behavior either in the workplace or on the phone, “misconduct” can indicate poor behavior and/or actions in the workplace, and “attendance” is for habitual absenteeism (which contrasts with “no show” which is when the agent simply stops coming to work, typically determined after 3-6 days without the agent coming to work).
9. <month>\_REVENUE\_GENERATED quantifies the actual amount of debt the agent has collected for the month. While a “promise to pay” is considered a positive outcome for a call, it does not necessarily indicate that a payment was received; the revenue\_generated figure does. The collection agency has a system in place that tracks the promises garnered by each agent and follows up on the account to see if a payment was received, and only then does the revenue get counted. If a call is “transferred to closer”, it is unknown whether the payment taken during that call counts toward the transferring agent or the transferred agent. NOTE that the call data almost always lists the original agent who took the call, not the agent the call was transferred to (though errors do occur). As far as I know, no strict debt collection quota exists.
10. <month>\_COMMISSION is a very complicated calculation that takes into account generated revenue, pay rate, generated revenue for the agent’s associated group, performance among peers, fines and penalties assessed as a result of an agent’s noncompliance with debt collection rules and regulations, among other factors. The explicit calculation should not be necessary. It should suffice to know that an agent’s commission is an assessment of the agent’s overall monthly performance, separate from his/her monthly generated revenue, but not an entirely independent. An agent may be ineligible for receiving a commission if his/her noncompliance fines exceed a certain amount, but this is uncommon.
11. <month>\_GROUP is the group an agent was assigned to for that month, and is provided as a means of indicating that the agent may have changed groups since his/her initial hire (which is indicated by src1 & src2). The values are meant to align with src1 & src2’s values, but some abbreviations differ; for example, “dtv” is DirecTV and “vz” is Verizon. The actual meaning for a month’s group can be inferred by the value of src1/src2 with which it is most in common. For example, if one looks at the agents who worked in group “ptm” in August, one can see that clearly this corresponds to the T-Mobile group, and it also implies that “g3a” corresponds to the Telecom group.
12. Only the months of July through December were provided because those were the only months given to us. Previous months could have been provided, but the available audio only goes back to mid-June, so July seemed a logical starting point. In general, a larger block of calendar time is desired to try to eliminate potential seasonality effects. For example, debt collection is greatly influenced by tax season; unfortunately, the beginning of 2015 was unavailable.
13. TARGET\_VALUE in the feature data is the suggested value given to that audio file which indicates its type of expected outcome for further predictions, otherwise known as the “prediction class”, “dependent variable”, or “experimental outcome”; see <https://en.wikipedia.org/wiki/Classification_rule>. These target values are a suggestion. Recall that the objective here is to predict agents that are at risk for termination, specifically voluntary termination without notice or by abandonment (“no show”). The audio files with a target value of 1 correspond to agents considered to be long-standing employees, and thus a “positive” predictive outcome for this objective. Likewise, audio files with a target value of 0 correspond to agents that have abandoned the job or resigned without notice. Audio files with no target value suggest that the call was associated with an agent that was not a good representative for either prediction class. Again, the values are only a suggestion, and as part of the project a team is allowed to reassign audio files to different prediction classes.
14. FEATURE\_VALUE<n> are the individual features calculated entirely from the audio signal itself. In other words, no data from either the agent data or call data are incorporated into the calculations. The calculations are proprietary and will not be divulged, but are based on signal processing techniques involving Fourier Analysis. Further, almost all of the features involve some assessment of the predicted emotions a speaker was experiencing over the course of a call. As with other parts of this project, the calculations can be subject to scrutiny. Therefore, it is entirely acceptable (and perhaps a good avenue of research) to see if some mathematical recombination of features can produce a single feature that is a better predictor than if the original ones were considered separately. Assessing the set of 176 features for e.g. multicollinearities, normality, ANCOVA parameters, etc. is also recommended.
15. SKILL NAME corresponds to the group values in the agent data. As with the monthly group values, the SKILL NAMEs are not a verbatim association with group, but can be inferred easily. Some values such as “General” and “JCAP” have no available explanation however. The letters after the underscore indicate the means in which the call was initiated. “HCI” indicates that the call was initiated through an outbound autodialer (if/when a debtor stays on the line or interacts with the IVR, the call gets automatically transferred to an agent). “Manual” is an outbound call manually dialed by the agent. Other values have not been given an explanation. Through my own experience, “QC” likely has something to do with Quality Control and “RPC” may indicate a pre-confirmed Right Party Contact.
16. CALL DIRECTION can be left blank, but is an Outbound call regardless. The omission is due to older data sources not providing the information.
17. It is possible for an agent to be re-hired; if that is the case, they will receive the same agent\_id. In other words, no individual has 2 or more agent\_ids.
18. Average tenure/time-to-termination of agents for the company can be estimated by looking at the time elapsed between hire date and termination date. To give another perspective, the company’s agent annual attrition rate is nearly 100%; that is, given the average number of agents the company keeps on payroll at any given time, the company will have hired twice that many over the course of a year.
19. I don’t know what is meant by “different equations for voluntary vs involuntary”. We’re simply looking for predictive performance on agent termination. This can include both types of termination, but voluntary termination is the type that the company can’t control so It is of more interest to predict.
20. The question of whether or not “call performance” (in whatever form it is measured) can be assessed (i.e. predicted) based off of the feature vectors is one objective of the project – it’s up to you to answer it!
21. Agent data is available for agents that did not terminate; they are indicated by the fact that they do not have a term\_date specified. The prediction criterion is not entirely specific. We merely seek to know if some measure of predictive performance can be seen for “agent attrition”: one objective of the project is to try and find the conditions or subset of termination criteria which can maximize predictive performance (if of course good predictive performance cannot be done simply for all terminated agents vs. agents who didn’t terminate).
22. CALL DURATION is provided for all calls, regardless of whether or not CALL END TIME is provided (which may be omitted because it wasn’t available for earlier data). Note that the CALL DURATION can differ from CALL END TIME minus CALL START TIME; this is due to physical telephony effects. CALL DURATION should be used instead.
23. Regarding source reliability, source 1 and source 2 should be treated as equally reliable, but the monthly data overrules either should a conflict arise, but only for that month. For example, if an agent’s group is “DISH” for source1 and “DIRECTV” for source 2, it is equally possible that the agent was assigned to either. If however Aug\_group for the same agent was “DSH”, then one can trust that the agent indeed was taking calls for DISH Network during the month of August. For a better determination of what group the agent worked for in the other months, refer to those months’ group value. The ultimate determination of the group in which the agent was taking calls however is in the call data’s SKILL NAME. If an agent’s group value in any given month is “#N/A”, this usually indicates that the agent was not taking collections calls, but was working in another department. An agent who is present in the call data but which is “not applicable” for a group assignment is usually a manager or a quality assurance representative. “#N/A” therefore has a different interpretation than if the column were simply blank. Blank values most likely indicate that the data was unavailable, in which case you would refer back to source1 or source2’s group for that agent for that month.

EVENING BATCH

1. If agents are disregarded in the analysis, it could potentially have an effect on the outcome, but one of the objectives of the project is to determine if such culling can improve predictive performance so doing so should not be ruled out.
2. Regarding Question 2, see morning answer 22. If FILESIZE is blank, it means the data was unavailable for that entry, not that the file was 0 KB in size (there are 13 entries that explicitly have 0 for the size, and these are anomalous). The main indicator of file size/availability (and therefore usefulness) is the existence of a feature vector. If a call has a feature vector, the audio file was available. The converse is not necessarily true: there exist calls that had an available audio file, but features were not extracted from it.
3. Regarding Question 3-5 & 10, 22, …, see morning answers 13 & 14.
4. Call data was provided on a daily basis and thus were combined into a single file. During the course of data logging, the company changed the format of the data to include extra pieces of information (call end time, skill name, call direction, file size).
5. Regarding Question 7, use your best judgement.
6. Regarding Question 8, 9 & 41, I need more explanation of what is requested. Perhaps we should schedule Skype time. When assessing model performance, a baseline model isn’t necessarily needed. True/False Positives/Negatives are determined based on the classification system, where “Positive” can be defined as a reliable agent and “Negative” can be an agent that resigned, for example. Further, creating a few models and highlighting their differences is a very good idea.
7. Regarding Question 11, the predictive model need not be “logistic” i.e. have categorical dependent variable values (in the parlance of statistical classification), nor does it have to be a binary classifier – consider multiclass, multi-label, and/or fuzzy classifiers. Exploration of the nature of the model itself is recommended.
8. Regarding Question 12, 38 & 47, see morning answer 23.
9. Best judgement of the group must be used to determine how to handle any data outliers or discrepancies that may exist, such as that found in Dec\_hourly\_rate or mismatching hire\_date\_src1 & src2.
10. The definitive indicator that an agent no longer works for the company is the term\_date. If the term\_date (for either source) is not blank, the agent was terminated. It is entirely possible that terminated agents may not have a specification for termination reason etc.
11. Regarding Question 16, 32 & 49, see morning answer 1.
12. Regarding Question 17, see morning answer 16.
13. ACCOUNT in the call data is the identifier of the debtor that was called. Quite frequently more than 1 call is placed to a debtor, so multiple instances are likely.
14. Regarding Question 19, 29 & 50, see morning answer 15.
15. Anyone that makes any assumption does so at his/her own risk.
16. Audio files were recorded at the same bitrate and are in stereo.
17. For questions regarding src1 & src2, see morning answers 2-6.
18. Regarding Question 25, see morning answer 4.
19. Regarding Question 27, 28, 43 & 46, see morning answers 6 & 10.
20. Note that the subject of the desired predictions, the agents, all have multiple calls associated with them. It is one objective of the project to determine which calls to use as training/cross validation data points and/or using what means of aggregation (if any) to combine multiple calls into a single data point.
21. In the CALL DURATION(HMS) label, HMS refers to Hours, Minutes, Seconds, but this is a misnomer; the call duration is in fact measured in seconds only.
22. Regarding Question 48, see morning answer 5.
23. Although each type of data (agent, call, feature) has a different time range of available values, it is probably best to consider analysis on the common timeframe among them (so July to December).
24. Regarding Question 52, see morning answer 11.
25. Termination type (voluntary/involuntary), like several other pieces of information among the call and agent data, need not be included in the analysis but is provided as possible avenues of exploration for improving predictive performance. Among the project’s objectives, being able to predict who terminates voluntarily versus involuntarily is not one of them.
26. It is one objective of the project to determine how such a predictive model might be applied in practice, such as whether to make predictions on a per-agent basis (and if so how to combine the agent’s calls into a prediction data point) or on a per-call basis.
27. There definitely is missing data in the agent data and in the call data. In terms of the columns, missing data is indicated by a blank value (although as previously discussed a blank value does not always indicate that it was missing). In terms of the rows, be aware that several agents given in the agent data take inbound calls, and inbound calls are not included in the call data. Moreover, the call data does not even represent all outbound calls that were placed in that time period for any given agent. This is an unfortunate consequence of the data that was available; it was not intentional. Likewise, there was call data that referenced agent\_ids that did not exist in the agent data sources; these have already been excluded from the given data set. Lastly, there were agents employed between July and December 2015 that are not included in the agent data. The agents left out were those either dedicated entirely to Spanish-speaking accounts, who had no HR data (source1 or source2), or who had no outbound calls available.