**Appendix for Norrick\_Chad\_Classification\_Project:**

Answers to questions as asked, regarding why customers churning.

Headers and numbering here match the convention as set forth in the project rubric.

**Classification Project**:

**Initial questions from requester**:

1. Could the month in which customers sign up influence churn? I.e. Do some customer cohort(s) (with cohort defined by tenure) churn at a higher rate?
   1. Yes, we noticed that the cohort of customers within their first 13 months did indeed churn at a higher rate than more established, or long term, customers. As such, this logic is illustrated in the first 5-6 cells in this notebook, where the two populations are split out, after we noticed this trend. (Plot of this churn on line chart not provided due to time constraint but was observed in other exploration with a bar chart, also shown in the exploration section.)
2. Are there features that indicate a higher propensity to churn? (like internet service, type of phone service, online security and backup, senior citizen category, paying more than x% of customers with the same services)
   1. Yes, we did observe several of these features that were more closely correlated to churn. Among them were:
      1. **internet service** (*specifically, those who have fiber connection);*
      2. **type of phone service** (*specifically, we did notice a subtle correlation with those users who had multiple phone lines, however, this product feature is often associated with users who have internet, namely fiber*);
      3. **contract type** (*a large proportion of the churn was associated with customers on month-to-month contracts*);
      4. and finally, a combination of **monthly charges, total charges and tenure** (*which are related to one another, as monthly charges and tenure drive total charges*).
   2. Relevant plots are attached in the notebook, in the exploration section.
3. Is there a price threshold for specific services, where the likelihood of churn increases once price for those services goes past that point? If so, at what point for what service(s)?
   1. Yes, this was observed. Specifically, we observed that phone customers (1 phone line) who also had internet AND streaming services churned when their monthly charges approached $70. We also observed that phone users (multiple lines) who had also had internet and streaming services churned when their monthly charges hit $75. The churn was more pronounced among users who had fiber and/or the full streaming package. Associated charts are attached in the exploration phase of this notebook.
4. If we looked at churn rate for each month-to-month customers AND 1-yr contract customers both after the 12-month point, are those rates comparable?
   1. No, the rates are not comparable. Specifically, at that point, we observed that the majority of the churn (with regard to contract types) was among month to month customers and somewhat for 1-year customers, but very little at this point for 2-year contracts. We surmised that 1-year customers churned when they were able to do so without penalty (after that initial 12-month period had ended).
   2. Also, we surmised that the reason for virtually no churn among 2-year contracts at this point, is likely due to customer unwillingness to pay a penalty for early cancellation of the contract before their initial commitment was reached. Finally, we observed that the churn among these three contact types was most pronounced among customers paying for Fiber, the fastest and most expensive internet service available. This is the point when we began to consider that very robust phone/internet packages commanded high monthly costs, and this high cost among many users was the driver to churn.

**Data Prep:**

1. Write a function that takes in a dataframe and computes and returns values by frequency for each column.
   1. Function included in the associated prepare.py file, accompanying this notebook.
2. Missing values:
   1. Only one field was found to have missing values. The total\_charges field contained missing values for 11 customers who had signed up but who had zero tenure as they were in their initial month of service. The decision to drop these observations from the data set was made for two reasons:
      1. This was such a small number of blanks, that removing these records would be statistically insignificant and not impact any predictions of churn, and,
      2. These customers in their first month of service don’t actually pertain to our mission of creating that model to predict churn. These observations were then dropped from the data set in the initial step in the preparation python script. Again, this was the only feature containing blanks/nulls.
3. Transformation/encoding of variables:
   1. The decision was made to encode any non-numeric/string/object field into a numeric field to enable these fields to be considered in the research/analysis/modeling. That said, any field that contained string-based values was encoded to numeric during the data prep stage via the prepare.py python script.
4. Scaling:
   1. Per the instructions in the project rubric, the total and monthly charges variables were scaled to better enable comparison with other variables.

**Data Exploration:**

1. Could month signed-up influence churn?
   1. See # 1 above in initial questions
2. Features that indicate higher propensity for churn?
   1. See # 2 above in initial questions
3. Price thresholds for churn, specifically related to various services?
   1. See # 3 above in initial questions
4. Churn after 12 months, of different contract types?
   1. See #4 above initial questions
5. Controlling for services, is the mean monthly charge different for churn vs not churn?
   1. Not answered due to time constraint.
6. How much of monthly charges can be explained by internet service type?
   1. Answer: 82% as shown in the exploration section. We hypothesized this number would be relatively high, as was noted in the exploration section for this item, “Since monthly charges are driven by telco-based services, i.e. phone, internet and other services, it stands to reason that internet services will be highly correlated to the monthly charges.”
7. Same question as the previous question, but also include phone services.
   1. Answer: 78%, as shown in the exploration section. We also hypothesized this number would be relatively high but not as high as the mere internet services alone, as was noted in the exploration section for this item, “Same as above, except that adding in the additional variable of phone line services, reduces the percent of monthly charges that can be explained by the two fields combined.”
8. Visualizations: comparing chosen independent variables against one another and against the dependent variable of churn (in this case, “churn\_encoded”). Relevant plots are in the exploration section. Some visualizations used in early exploration, but most of the visualizations included here were performed AFTER the most correlated fields were discovered in that early exploration phase, and then those fields were plotted going forward. Again, plots displayed in exploration section of this notebook.
   1. Plots not used due to the time constraint: pd.scatter\_matrix, sns.swarmplot, and sns.boxplot (although a series.plot.box was used instead)
   2. Use at least one more type of plot not included in the list above: We used a pair grid in the initial exploration, combined with plt.hist and plt.scatter.
9. What can you say about each variable’s relationship to churn, based on initial exploration? If there appears to be some sort of interaction or correlation, assume there is no causal relationship and brainstorm (and document) ideas on reasons there could be correlation.
   1. The variables chosen after exploration include internet service type, contract type, streaming services, tenure, and both the monthly and total charges. However, after further consideration, the decision was made to remove the monthly and total charges fields, as these fields are driven by a combination of tenure and the various services the customer consumes. They are addressed below:
   2. Internet service type was very closely correlated to each of our populations’ rate of churn. This is very likely due to the increasingly more expensive internet packages, proving greater bandwidth but at an increased cost.
   3. Contract type is also related to churn for both populations, but for differing reasons. Among the month to month customers, there is no customer penalty for churning. Hence, it is financially easier for these customers to churn, if they feel their monthly charges are too high for the services provided. Among long term customers there IS a penalty for churning, resulting in an inverse correlation to churn among these particular customers. This is the probable reason we did not observe any significant churn among 1-year and 2-year-contract customers, UNTIL they hit the 13th and 25th months, respectively.
   4. For customers less than one-year, streaming services is also related to churn. I believe this is similar to fiber in the internet services field. Streaming and fiber are both costly add-ons to a telco package of services. As such, I believe the piling on of these additional services drives a customer’s monthly charges up, which leads to churn.
   5. For long term customers, tenure was associated with churn. However, I believe this to be related to the interlocking relationship among tenure/services/charges, where a combination of these drives the other. I believe tenure is only correlated because of its relationship to charges, which is also related to services. Services (and the various associated costs) appears to be a large driver in churn.
   6. See exploration section for all relevant plots supporting these hunches.
10. Summarize your conclusions, provide clear answers and summarize any takeaways/action plan for the work above.
    1. As was stated in the section above, I believe our initial peek at these data show relationships driving churn, specifically related to costs, driven by additional services or premium services.
    2. Definitely, further research needs to be done on this data set, as we simply ran out of time to dive deeper and build the pieces necessary to show specifically these relationships/correlations in greater detail.
    3. At this point, I believe we have provided the 80% solution on these data, given the time constraint. Given additional time, I do believe we would be able achieve the 100% solution.
    4. Possible recommendations: our research leads us to believe that the greatest chunk of churn is among month to month customers within a first year of service. Specifically, we believe these customers to be consumers of larger packages of services, namely fiber and streaming, among others. A possible recommendation would be to provide these services at a discounted rate for any customer in their first year, or even second year of tenure. Providing discounts may ultimately be less costly than replacing customers who churn.
    5. FINAL TAKEAWAY: As was mentioned in the rubric, customers are leaving not because they can, but more likely for another reason, because they are motivated to do so. We hypothesize they are motivated to churn because they believe they can find the same services cheaper at a competitor. As older, more loyal customers slowly fall from this market, they are replaced by consumers more willing to purchase more robust telco packages of services. If we cannot provide those services to them at a price they’re willing to pay, they will simply look elsewhere.