Customer Churn

Springboard Capstone 2 Milestone Report 1

Brock Nosbisch

10.11.2019

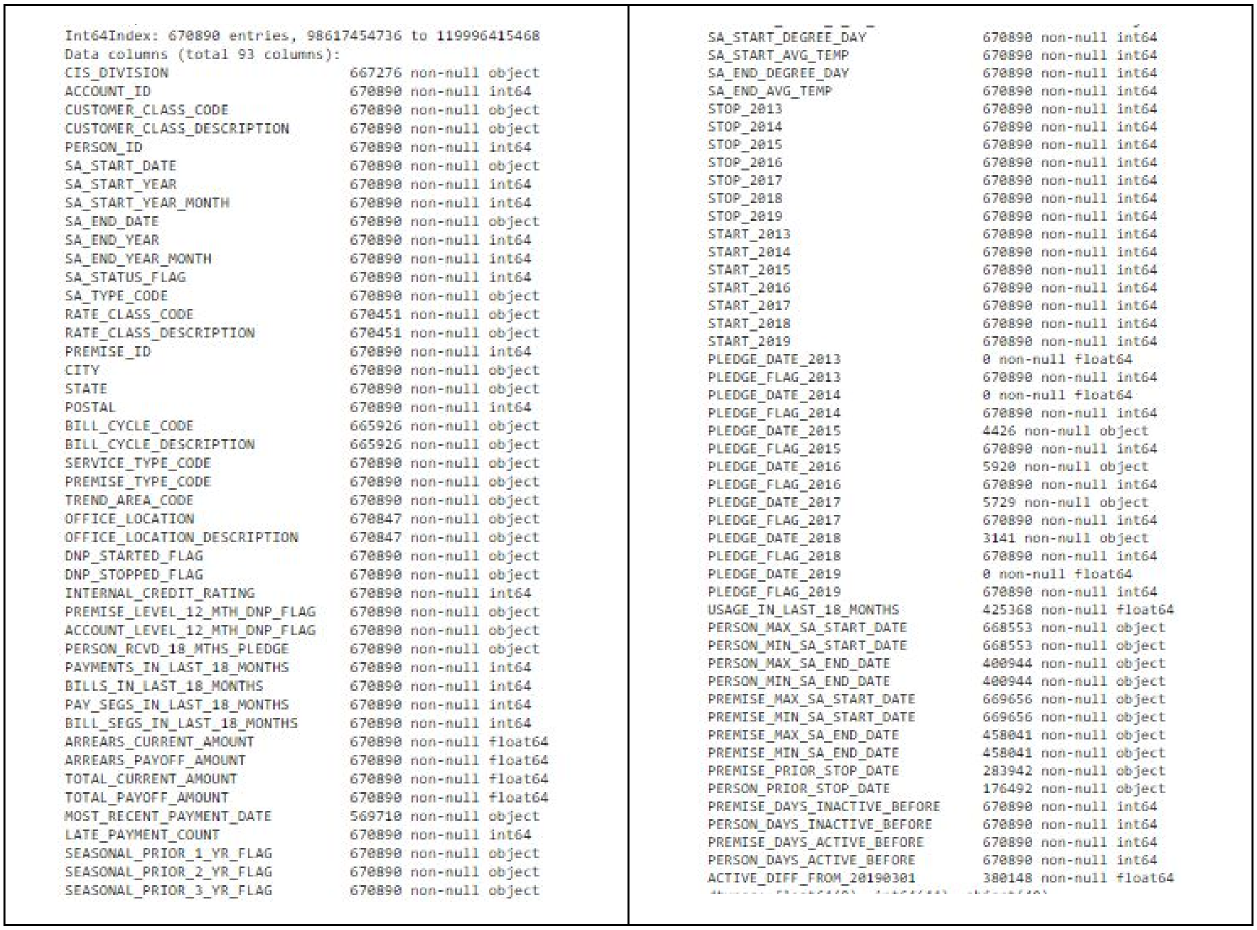
Problem Statement

My client is a natural gas company. Being able to predict when a customer will turn off is crucial to the natural gas business. By knowing when a customer will turn off, you can forecast future turn off work and also allow sales reps to reach out to those customers to try to get them to stay on.

The Dataset

The dataset used in this project comes from both internal data and external data sources. The data will be merged inside our internal data warehouse and then the dataset will be manually downloaded and imported into the Data Wrangling notebook.

The original data is sourced from multiple applications and external sources. It is combined in a data warehouse and then exported with the below columns. The imported data set contains 670,000 records with 93 columns.



```
Int64Index: 670890 entries, 98617454736 to 119996415468
Data columns (total 93 columns):
CIS_DIVISION                    667276 non-null object
ACCOUNT_ID                      670890 non-null int64
CUSTOMER_CLASS_CODE             670890 non-null object
CUSTOMER_CLASS_DESCRIPTION      670890 non-null object
PERSON_ID                       670890 non-null int64
SA_START_DATE                   670890 non-null object
SA_START_YEAR                   670890 non-null int64
SA_START_YEAR_MONTH             670890 non-null int64
SA_END_DATE                     670890 non-null object
SA_END_YEAR                     670890 non-null int64
SA_END_YEAR_MONTH               670890 non-null int64
SA_STATUS_FLAG                  670890 non-null int64
SA_TYPE_CODE                    670890 non-null object
RATE_CLASS_CODE                 670451 non-null object
RATE_CLASS_DESCRIPTION          670451 non-null object
PREMISE_ID                      670890 non-null int64
CITY                            670890 non-null object
STATE                           670890 non-null object
POSTAL                          670890 non-null int64
BILL_CYCLE_CODE                 665926 non-null object
BILL_CYCLE_DESCRIPTION          665926 non-null object
SERVICE_TYPE_CODE               670890 non-null object
PREMISE_TYPE_CODE               670890 non-null object
TREND_AREA_CODE                 670890 non-null object
OFFICE_LOCATION                 670847 non-null object
OFFICE_LOCATION_DESCRIPTION     670847 non-null object
DNP_STARTED_FLAG                670890 non-null object
DNP_STOPPED_FLAG                670890 non-null object
INTERNAL_CREDIT_RATING          670890 non-null int64
PREMISE_LEVEL_12_MTH_DNP_FLAG   670890 non-null object
ACCOUNT_LEVEL_12_MTH_DNP_FLAG   670890 non-null object
PERSON_RCVD_18_MTHS_PLEDGE      670890 non-null object
PAYMENTS_IN_LAST_18_MONTHS      670890 non-null int64
BILLS_IN_LAST_18_MONTHS         670890 non-null int64
PAY_SEGS_IN_LAST_18_MONTHS      670890 non-null int64
BILL_SEGS_IN_LAST_18_MONTHS     670890 non-null int64
ARREARS_CURRENT_AMOUNT          670890 non-null float64
ARREARS_PAYOFF_AMOUNT           670890 non-null float64
TOTAL_CURRENT_AMOUNT            670890 non-null float64
TOTAL_PAYOFF_AMOUNT             670890 non-null float64
MOST_RECENT_PAYMENT_DATE        569710 non-null object
LATE_PAYMENT_COUNT              670890 non-null int64
SEASONAL_PRIOR_1_YR_FLAG        670890 non-null object
SEASONAL_PRIOR_2_YR_FLAG        670890 non-null object
SEASONAL_PRIOR_3_YR_FLAG        670890 non-null object
```

```
SA_START_DEGREE_DAY             670890 non-null int64
SA_START_AVG_TEMP               670890 non-null int64
SA_END_DEGREE_DAY               670890 non-null int64
SA_END_AVG_TEMP                 670890 non-null int64
STOP_2013                       670890 non-null int64
STOP_2014                       670890 non-null int64
STOP_2015                       670890 non-null int64
STOP_2016                       670890 non-null int64
STOP_2017                       670890 non-null int64
STOP_2018                       670890 non-null int64
STOP_2019                       670890 non-null int64
START_2013                      670890 non-null int64
START_2014                      670890 non-null int64
START_2015                      670890 non-null int64
START_2016                      670890 non-null int64
START_2017                      670890 non-null int64
START_2018                      670890 non-null int64
START_2019                      670890 non-null int64
PLEDGE_DATE_2013                0 non-null float64
PLEDGE_FLAG_2013                670890 non-null int64
PLEDGE_DATE_2014                0 non-null float64
PLEDGE_FLAG_2014                670890 non-null int64
PLEDGE_DATE_2015                4426 non-null object
PLEDGE_FLAG_2015                670890 non-null int64
PLEDGE_DATE_2016                5920 non-null object
PLEDGE_FLAG_2016                670890 non-null int64
PLEDGE_DATE_2017                5729 non-null object
PLEDGE_FLAG_2017                670890 non-null int64
PLEDGE_DATE_2018                3141 non-null object
PLEDGE_FLAG_2018                670890 non-null int64
PLEDGE_DATE_2019                0 non-null float64
PLEDGE_FLAG_2019                670890 non-null int64
USAGE_IN_LAST_18_MONTHS         425368 non-null float64
PERSON_MAX_SA_START_DATE        668553 non-null object
PERSON_MIN_SA_START_DATE        668553 non-null object
PERSON_MAX_SA_END_DATE          400944 non-null object
PERSON_MIN_SA_END_DATE          400944 non-null object
PREMISE_MAX_SA_START_DATE       669656 non-null object
PREMISE_MIN_SA_START_DATE       669656 non-null object
PREMISE_MAX_SA_END_DATE         458041 non-null object
PREMISE_MIN_SA_END_DATE         458041 non-null object
PREMISE_PRIOR_STOP_DATE         283942 non-null object
PERSON_PRIOR_STOP_DATE          176492 non-null object
PREMISE_DAYS_INACTIVE_BEFORE    670890 non-null int64
PERSON_DAYS_INACTIVE_BEFORE     670890 non-null int64
PREMISE_DAYS_ACTIVE_BEFORE      670890 non-null int64
PERSON_DAYS_ACTIVE_BEFORE       670890 non-null int64
ACTIVE_DIFF_FROM_20190301       380148 non-null float64
```

Data Wrangling

As mentioned, most of the data manipulation and joins were done in SQL prior to loading the data into the Python notebook.

The dataset contains Service Agreement (gas service) level data for random cities. The aggregated data is as of 3/1/2019 for this project's purpose. This will later be updated so that Python will dynamically generate the data being brought in. For now the data is generated from the "Queries - Used in Data Wrangling.sql" file in this repository. Some code was removed for masking and PII concerns. There are several temp tables created in the mentioned .sql file. At a high level, the data contains:

- Service Agreement level data
- Weather information
- Turn Ons
- Turn Offs
- Credit Scores
- DNP flags
- Number of pledges, payments, bills, late payments, and usage in the past 18 months
- Total money owed
- Prior 3 years of being a seasonal customer
- Prior premise and person level stops/starts.

The data will be read into the Python notebook through a csv (data.csv).

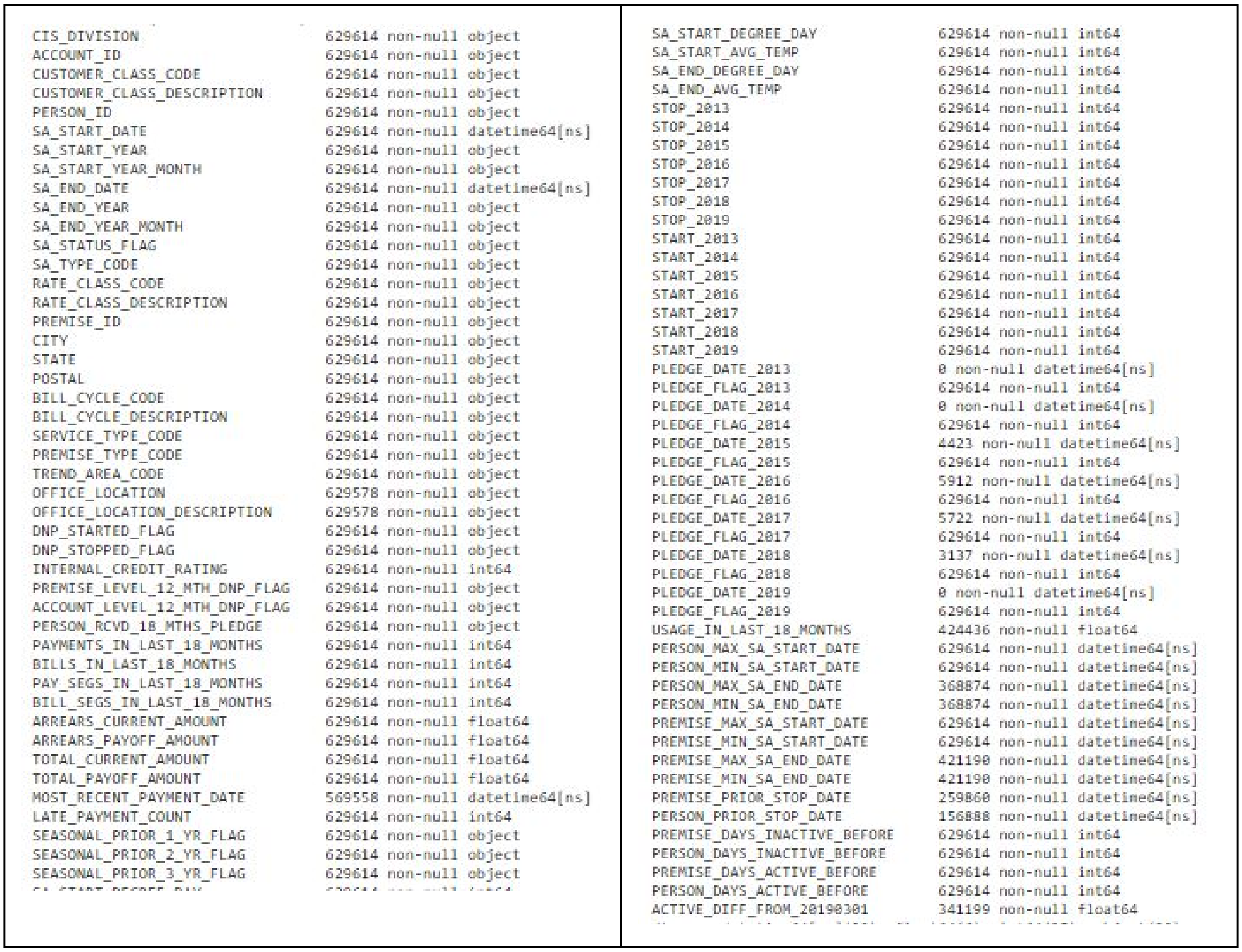
I will then set the index for the dataset to the unique identifier.

There are several fields that will need to have their data types updated and I will remove all Pending Starts, Reactivated, and Cancelled SAs since these should not be considered for Stops.

I will also remove a few bad records that have Null Rate Class, Company, and Bill Cycle.

Updated Dataset After Data Wrangling

For the most part, the columns are the same as what was read into the notebook. The big difference is in the data types and removal of some of the records.



```
CIS_DIVISION                    629614 non-null object
ACCOUNT_ID                      629614 non-null object
CUSTOMER_CLASS_CODE             629614 non-null object
CUSTOMER_CLASS_DESCRIPTION      629614 non-null object
PERSON_ID                       629614 non-null object
SA_START_DATE                   629614 non-null datetime64[ns]
SA_START_YEAR                   629614 non-null object
SA_START_YEAR_MONTH             629614 non-null object
SA_END_DATE                     629614 non-null datetime64[ns]
SA_END_YEAR                     629614 non-null object
SA_END_YEAR_MONTH               629614 non-null object
SA_STATUS_FLAG                  629614 non-null object
SA_TYPE_CODE                    629614 non-null object
RATE_CLASS_CODE                 629614 non-null object
RATE_CLASS_DESCRIPTION          629614 non-null object
PREMISE_ID                      629614 non-null object
CITY                            629614 non-null object
STATE                           629614 non-null object
POSTAL                          629614 non-null object
BILL_CYCLE_CODE                 629614 non-null object
BILL_CYCLE_DESCRIPTION          629614 non-null object
SERVICE_TYPE_CODE               629614 non-null object
PREMISE_TYPE_CODE               629614 non-null object
TREND_AREA_CODE                 629614 non-null object
OFFICE_LOCATION                 629578 non-null object
OFFICE_LOCATION_DESCRIPTION     629578 non-null object
DNP_STARTED_FLAG                629614 non-null object
DNP_STOPPED_FLAG                629614 non-null object
INTERNAL_CREDIT_RATING          629614 non-null int64
PREMISE_LEVEL_12_MTH_DNP_FLAG   629614 non-null object
ACCOUNT_LEVEL_12_MTH_DNP_FLAG   629614 non-null object
PERSON_RCVD_18_MTHS_PLEDGE      629614 non-null object
PAYMENTS_IN_LAST_18_MONTHS      629614 non-null int64
BILLS_IN_LAST_18_MONTHS         629614 non-null int64
PAY_SEGS_IN_LAST_18_MONTHS      629614 non-null int64
BILL_SEGS_IN_LAST_18_MONTHS     629614 non-null int64
ARREARS_CURRENT_AMOUNT          629614 non-null float64
ARREARS_PAYOFF_AMOUNT           629614 non-null float64
TOTAL_CURRENT_AMOUNT            629614 non-null float64
TOTAL_PAYOFF_AMOUNT             629614 non-null float64
MOST_RECENT_PAYMENT_DATE        569558 non-null datetime64[ns]
LATE_PAYMENT_COUNT              629614 non-null int64
SEASONAL_PRIOR_1_YR_FLAG        629614 non-null object
SEASONAL_PRIOR_2_YR_FLAG        629614 non-null object
SEASONAL_PRIOR_3_YR_FLAG        629614 non-null object
```

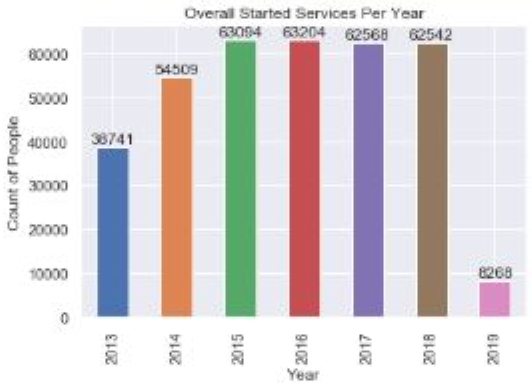
```
SA_START_DEGREE_DAY             629614 non-null int64
SA_START_AVG_TEMP               629614 non-null int64
SA_END_DEGREE_DAY               629614 non-null int64
SA_END_AVG_TEMP                 629614 non-null int64
STOP_2013                       629614 non-null int64
STOP_2014                       629614 non-null int64
STOP_2015                       629614 non-null int64
STOP_2016                       629614 non-null int64
STOP_2017                       629614 non-null int64
STOP_2018                       629614 non-null int64
STOP_2019                       629614 non-null int64
START_2013                      629614 non-null int64
START_2014                      629614 non-null int64
START_2015                      629614 non-null int64
START_2016                      629614 non-null int64
START_2017                      629614 non-null int64
START_2018                      629614 non-null int64
START_2019                      629614 non-null int64
PLEDGE_DATE_2013                0 non-null datetime64[ns]
PLEDGE_FLAG_2013                629614 non-null int64
PLEDGE_DATE_2014                0 non-null datetime64[ns]
PLEDGE_FLAG_2014                629614 non-null int64
PLEDGE_DATE_2015                4423 non-null datetime64[ns]
PLEDGE_FLAG_2015                629614 non-null int64
PLEDGE_DATE_2016                5912 non-null datetime64[ns]
PLEDGE_FLAG_2016                629614 non-null int64
PLEDGE_DATE_2017                5722 non-null datetime64[ns]
PLEDGE_FLAG_2017                629614 non-null int64
PLEDGE_DATE_2018                3137 non-null datetime64[ns]
PLEDGE_FLAG_2018                629614 non-null int64
PLEDGE_DATE_2019                0 non-null datetime64[ns]
PLEDGE_FLAG_2019                629614 non-null int64
USAGE_IN_LAST_18_MONTHS         424436 non-null float64
PERSON_MAX_SA_START_DATE        629614 non-null datetime64[ns]
PERSON_MIN_SA_START_DATE        629614 non-null datetime64[ns]
PERSON_MAX_SA_END_DATE          368874 non-null datetime64[ns]
PERSON_MIN_SA_END_DATE          368874 non-null datetime64[ns]
PREMISE_MAX_SA_START_DATE       629614 non-null datetime64[ns]
PREMISE_MIN_SA_START_DATE       629614 non-null datetime64[ns]
PREMISE_MAX_SA_END_DATE         421190 non-null datetime64[ns]
PREMISE_MIN_SA_END_DATE         421190 non-null datetime64[ns]
PREMISE_PRIOR_STOP_DATE         259860 non-null datetime64[ns]
PERSON_PRIOR_STOP_DATE          156888 non-null datetime64[ns]
PREMISE_DAYS_INACTIVE_BEFORE    629614 non-null int64
PERSON_DAYS_INACTIVE_BEFORE     629614 non-null int64
PREMISE_DAYS_ACTIVE_BEFORE      629614 non-null int64
PERSON_DAYS_ACTIVE_BEFORE       629614 non-null int64
ACTIVE_DIFF_FROM_20190301       341199 non-null float64
```

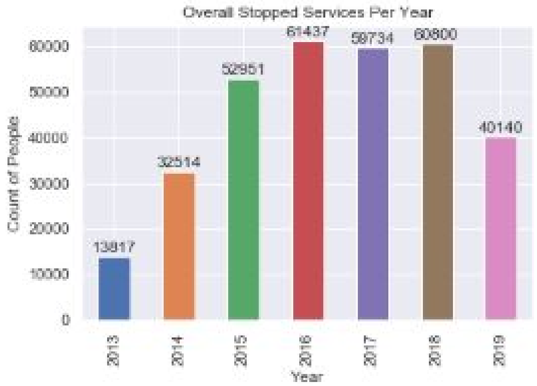
Field Definitions

- **CIS Division**: Company
- **Account ID**: An account is owned by 1 person but can have multiple Service Agreements.
- **Customer Class**: Account level customer class.
- **Person ID**: A person can have multiple accounts.
- **SA Start**: The Service Agreement's Start Date (when gas started).
- **SA End**: The Service Agreement's End Date (when gas stopped).
- **SA Status**: The current SA status.
- **SA Type**: The Service Agreement's type (example: Gas Residential).
- **Rate Class**: Determines how the SA is billed.
- **Premise ID**: Unique identifier for an address.
- **City**: Premise's City
- **State**: Premise's State
- **Postal**: Premise's Zip Code
- **Bill Cycle**: How we know when to bill an Account.
- **Service Type**: Always Gas.
- **Premise Type**: Premise Type (example: Commercial).

- **Office Location**: Closest office to premise.
- **DNP Started Flag**: Was the previous stop a DNP.
- **DNP Stopped Flag**: Did the current SA Stop because of DNP.
- **Internal Credit Rating**: The internal credit rating for an Account.
- **Premise Level 12 Month DNP Flag**: Has the Premise had a DNP in the last 12 months.
- **Account Level 12 Month DNP Flag**: Has the Account had a DNP in the last 12 months.
- **Person Received 18 Months Pledge**: Has the Person received a Pledge in the last 18 months.
- **Payments in Last 18 Months**: How many payments have been made for the SA in the last 18 months.
- **Bills in Last 18 Months**: How many bills were billed for the SA in the last 18 months.
- **Arrears Current/Payoff Amount**: How much money is past due for the SA.
- **Total Current/Payoff Amount**: How much money is owed for the SA.
- **Most Recent Payment Date**: What was the most recent Payment Date.
- **Late Payment Count**: How many late payments were charged to the SA in the past 18 months.
- **Seasonal Flags (1 year ago, 2 years ago, 3 years ago)**: Did the premise turn off between 3/1 and 6/30 and then turn on between 8/1 and 12/31.
- **SA Start/End Degree**: What was the Degree Days on the SA Start and SA End dates.
- **Stop/Start Flags**: Flags to show if the SA started/stopped in a given year.
- **Pledge Date/Flag**: The Pledge Date and Flag for each year.
- **Usage in Last 18 Months**: The Usage billed in the past 18 months.
- **Person/Premise Min/Max SA Start/End Dates**: What are the minimum and maximum start and end dates at a Person or Premise level.
- **Person/Premise Days Active/Inactive Before**: How many days was the previous SA for the Person or Premise Active or Inactive for.
- **Stop Days from 20190301**: Flags to use for training/testing. This should be dynamic but we are hardcoding the value for now.

Overall Started/Stopped Services Per Year

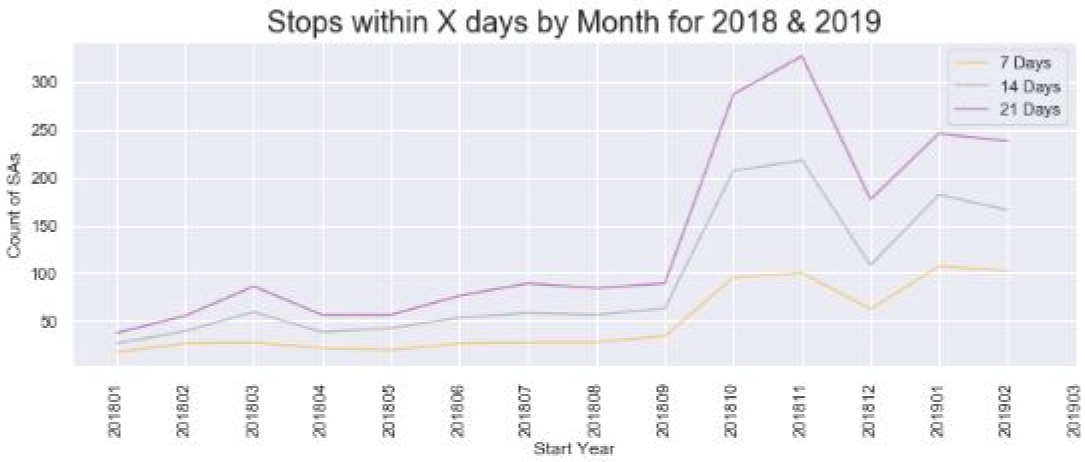




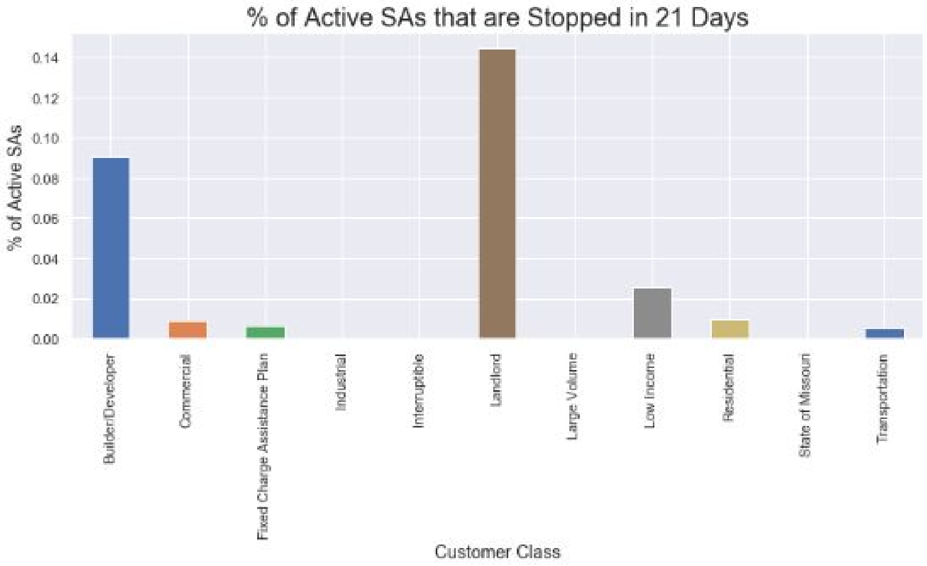
Exploratory Data Analysis

During EDA, several observations of interest were found.

1) The most frequent Start Date of the SAs that were stopped in March had started gas service in October, November, and December of the previous year.

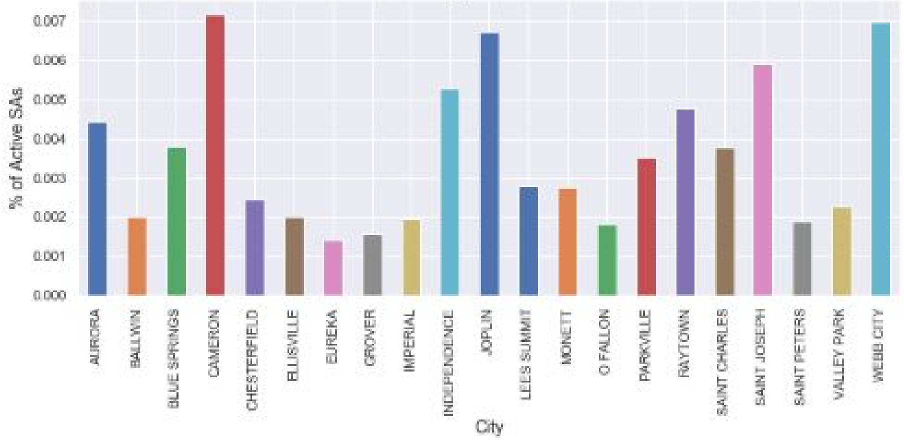


2) The Residential Customer Class has the most stops during all 3 of the time windows but less than 1% of all active Residential SAs were stopped. The biggest offenders of stops during the 3 windows are Landlords and Builder/Developers. 14% of Landlord SAs stopped during the 21 day window and 9% of Builder/Developer SAs stopped during the same window.

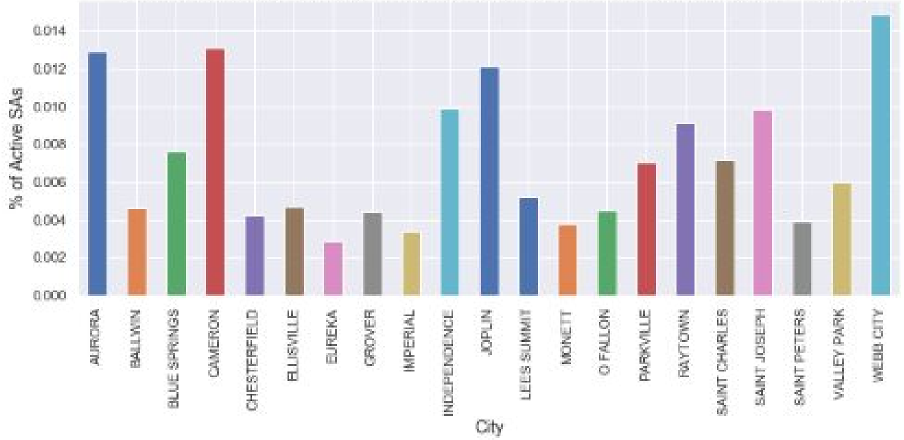


3) When looking at stops during the 7 day, 14 day, and 21 day windows by City, Webb City, Joplin, Cameron, and Aurora have the highest % of Active SAs that were stopped. Over 2% of all Active SAs in Aurora are stopped during the 21 day window.

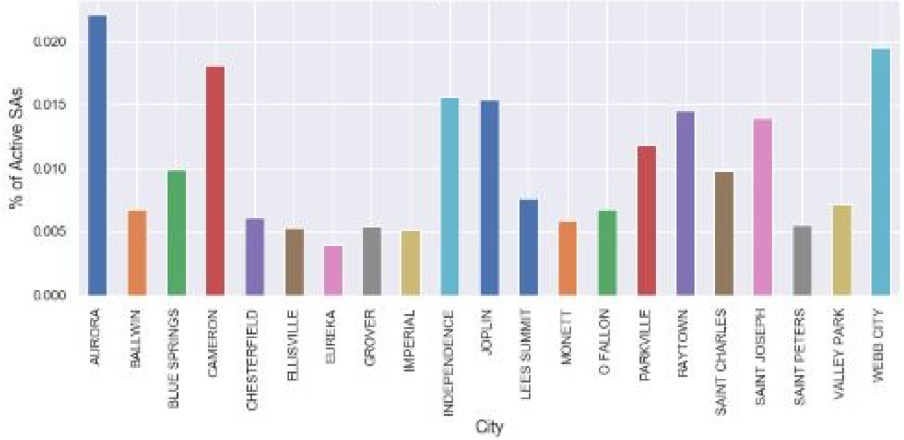
% of Active SAs that were Stopped between 3/1/2019 and 3/7/2019



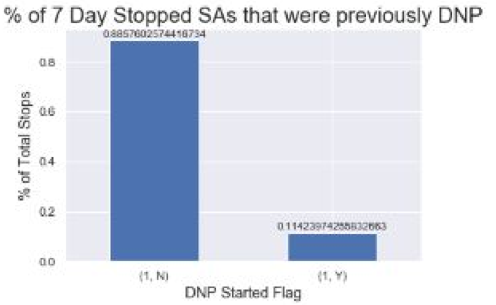
% of Active SAs that were Stopped between 3/1/2019 and 3/14/2019

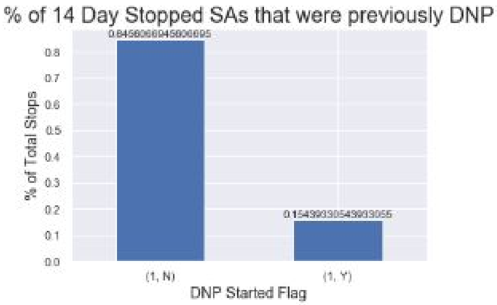


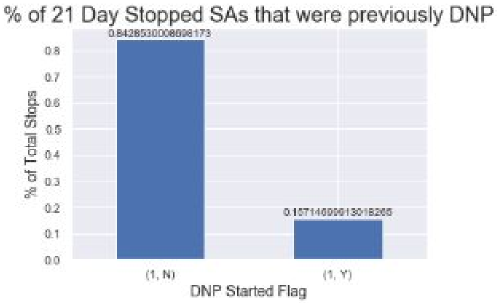
% of Active SAs that were Stopped between 3/1/2019 and 3/21/2019



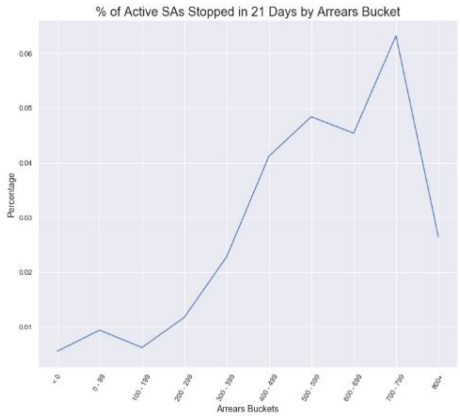
4) The percentage of stopped SAs that were DNPs previously during the 7 day window was 11%, 14 day window was 15%, and 21 day window was also 15%.

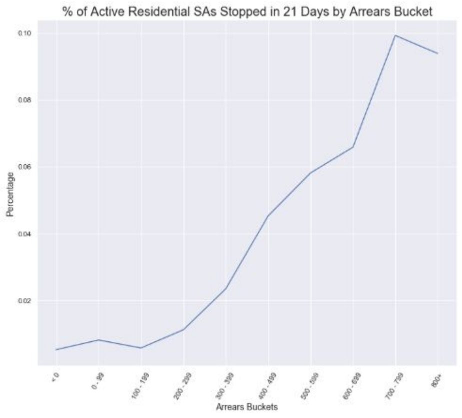




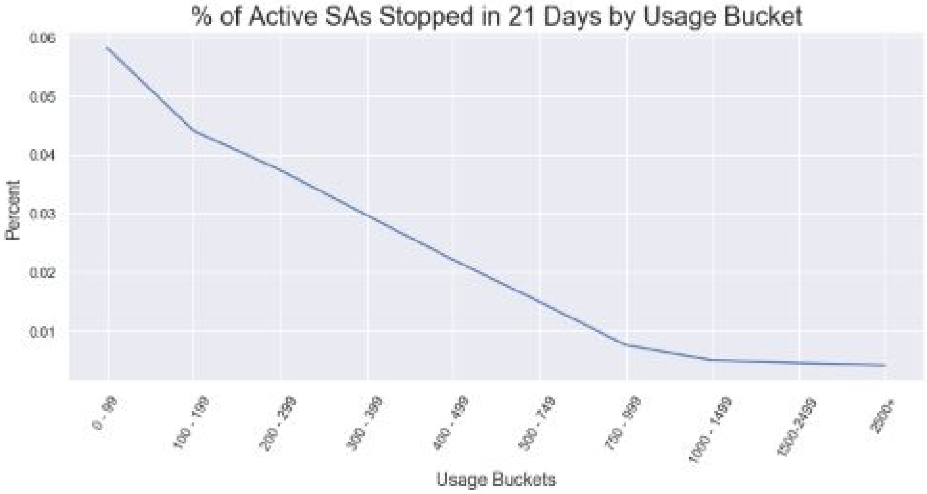


5) The higher the arrears amount, the more likely you are to stop. For example, if you owe 100 or less then you have less than a 1% chance of stopping between 3/1/19 - 3/21/19. If you owe 700−799 then you have over a 6% chance of stopping in the same timeframe. Residential customers have almost a 10% chance of stopping in the same bucket/timeframe.

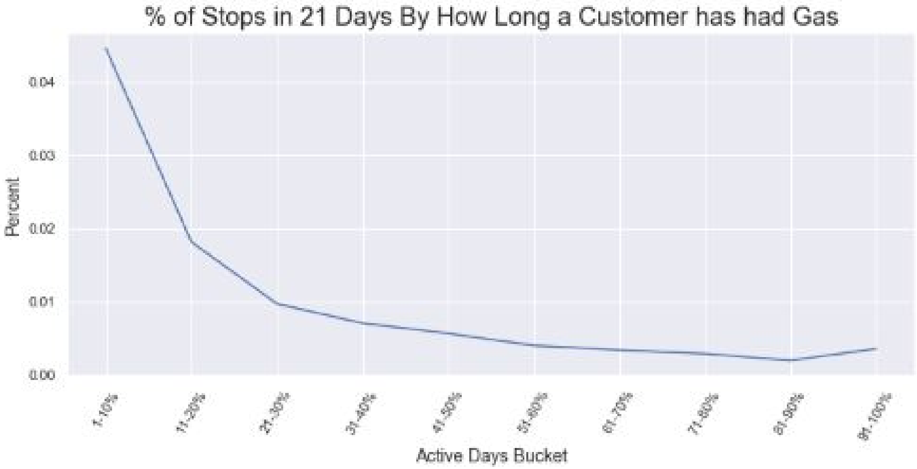




6) For Usage, it appears that the less you use, the higher chance you have of stopping service. Overall, if you have used < 100 therms then you have almost a 6% chance of stopping.



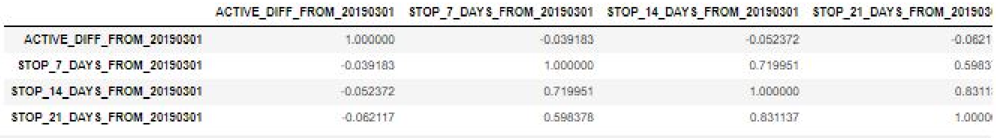
7) The shorter amount of time a customer has had gas service, the more likely the customer will stop gas.



Statistical Inference

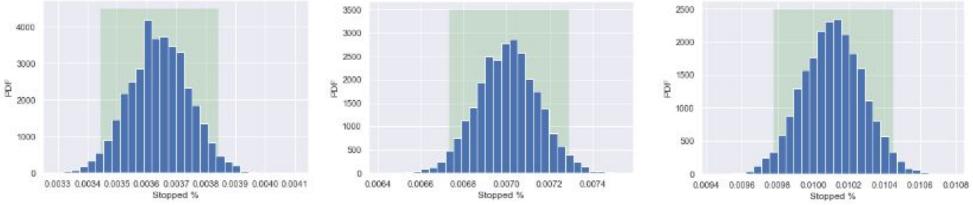
To confirm some of the above observations that may prove to be useful, I applied inferential statistics and found the correlations that are of statistical significance and a few that have practical significance. Here are a few observations:

- When comparing the Pearson correlation between Late Payment Counts and Stopping 21 Days from 3/1/2019, it has a p-value > .05 so the Null Hypothesis holds true. We can assume there is no significant correlation between these 2 values.
- # of Days Active as of 3/1/2019 has a negative correlation with Stopping. So in other words, the less days you have been active, the more likely you are to stop within 21 days of 3/1/2019.

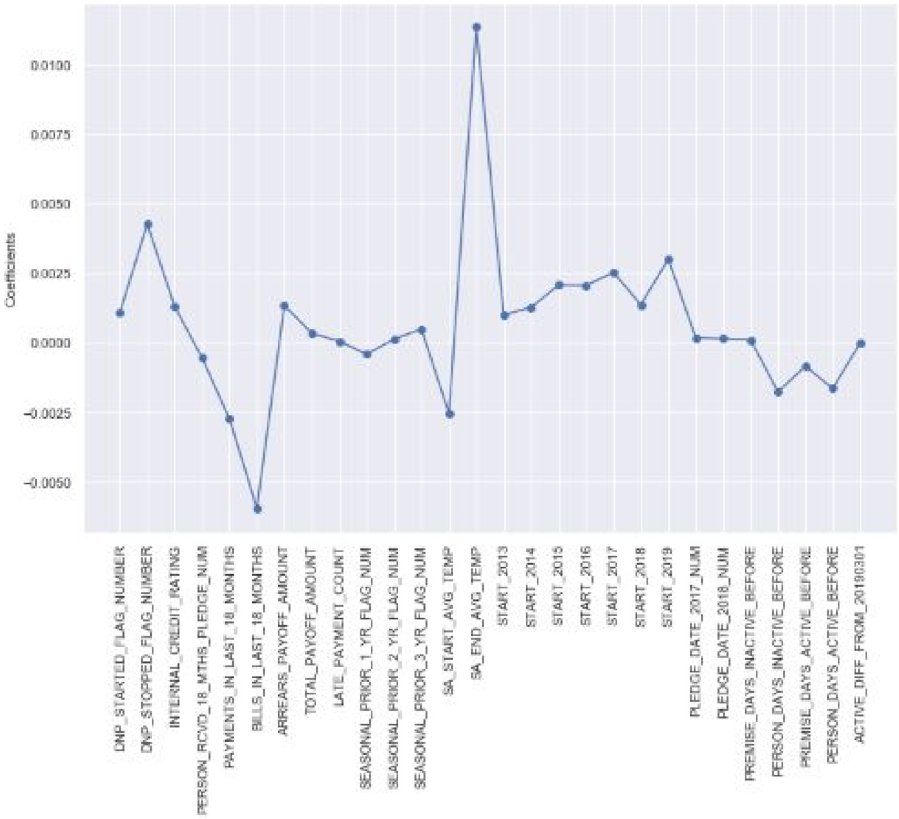


	ACTIVE_DIFF_FROM_20190301	STOP_7_DAYS_FROM_20190301	STOP_14_DAYS_FROM_20190301	STOP_21_DAYS_FROM_201903
ACTIVE_DIFF_FROM_20190301	1.000000	-0.039183	-0.052372	-0.0621
STOP_7_DAYS_FROM_20190301	-0.039183	1.000000	0.719951	0.5983
STOP_14_DAYS_FROM_20190301	-0.052372	0.719951	1.000000	0.8311
STOP_21_DAYS_FROM_20190301	-0.062117	0.598378	0.831137	1.0000

- The Bootstrap Estimate of SA Stops shows the 95% confidence interval for SAs stopping between 3/1/2017 and 3/7/2019 is from .34% to .38%.. Similarly, 3/1/2017 to 3/14/2019 is from .67% to .72% and 3/1/2019 to 3/21/2019 is from .98% to 1.04%.



- I used Lasso for Feature Selection. At a low value, the results show the Bills in last 18 months, DNP Stopped Flag, and SA End Avg Temperature as the top features.


Coefficients
DNP_STARTED_FLAG_NUMBER
DNP_STOPPED_FLAG_NUMBER
INTERNAL_CREDIT_RATING
PERSON_RCVD_18_MTHS_PLEDGE_NUM
PAYMENTS_IN_LAST_18_MONTHS
BILLS_IN_LAST_18_MONTHS
ARREARS_PAYOFF_AMOUNT
TOTAL_PAYOFF_AMOUNT
LATE_PAYMENT_COUNT
SEASONAL_PRIOR_1_YR_FLAG_NUM
SEASONAL_PRIOR_2_YR_FLAG_NUM
SEASONAL_PRIOR_3_YR_FLAG_NUM
SA_START_AVG_TEMP
SA_END_AVG_TEMP
START_2013
START_2014
START_2015
START_2016
START_2017
START_2018
START_2019
PLEDGE_DATE_2017_NUM
PLEDGE_DATE_2018_NUM
PREMISE_DAYS_INACTIVE_BEFORE
PERSON_DAYS_INACTIVE_BEFORE
PREMISE_DAYS_ACTIVE_BEFORE
PERSON_DAYS_ACTIVE_BEFORE
ACTIVE_DIFF_FROM_20190301