Customer Retention Analysis Small Business | Tax Industry

Miguel Santana

Flatiron School

Data Science | FT Cohort

Introduction

Methodology

OSEMN Framework

Obtain | Scrub

Exploratory Data Analysis

Modeling

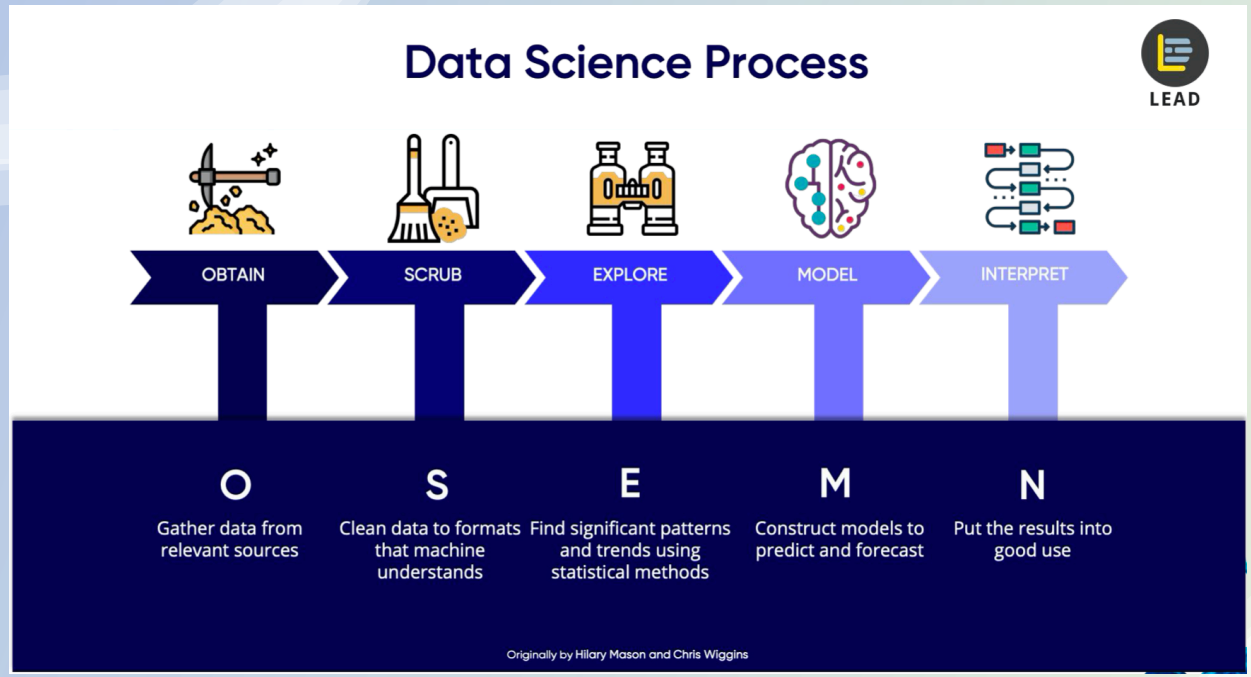
Analyzing Results

Business Recommendations

Future Work | Limitations

Methodology

- A small tax office is looking to understand their client base and prepare for the upcoming tax season. COVID-19 has severely effected the staff's ability to build rapport with clients, so our team was tasked with establishing trends in customer churn to be used customer retentive activities including targeted marketing.
- Framework: OSEMN
  - Merging client data spanning over a three years
  - Differentiating between current, returning, new and lost customers
  - Measuring customer churn while excluding new customers
  - Building Models and Analyzing the results



OSEMN Framework

Obtain | Scrub

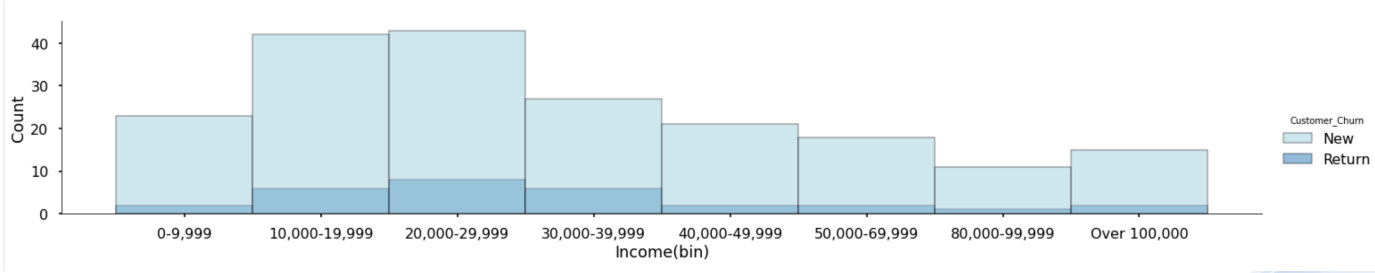
- Anonymizing the data
- Our focus was on E-Filing customers (per the owner's request)
  - Removed clients with no E-File Date
- Merging files (Tax Years 2016, 2017, 2018)
- Starting with most recent info and filling in missing values with data from previous years.
  - Preventing any data loss due to changes in demographic info
- Converting E-File dates to time series for analysis and feature engineering.
- Identifying current, returning, new and lost customers.

Exploratory Data Analysis

New Customers vs. Returning Customers

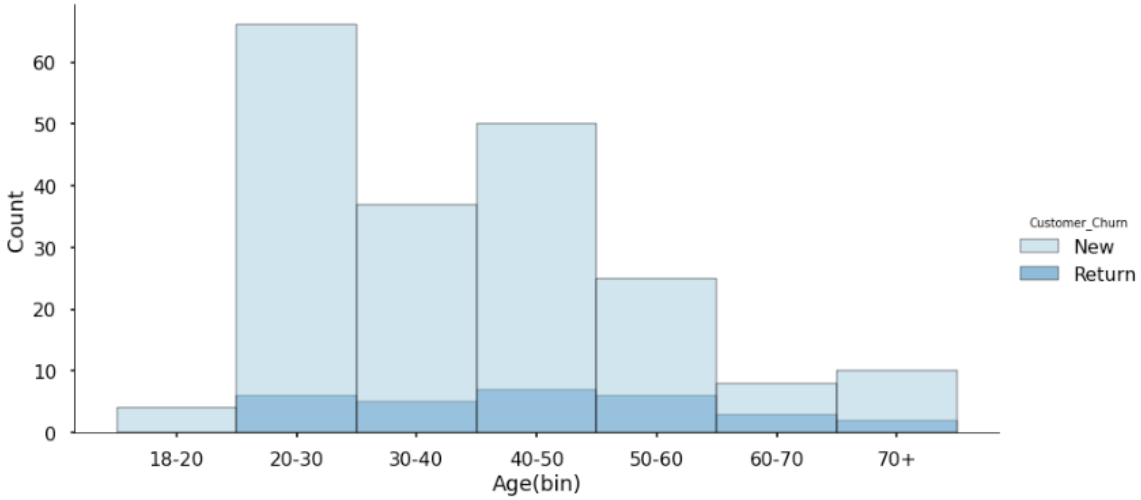
Taxable Income

- New and Return customers follow similar taxable income trends.
- $10,000 - $30,000



Age

- New Customers
  - Majority: 20-50 years old
- Return Customers
  - Majority: 40-60 years old



Modeling

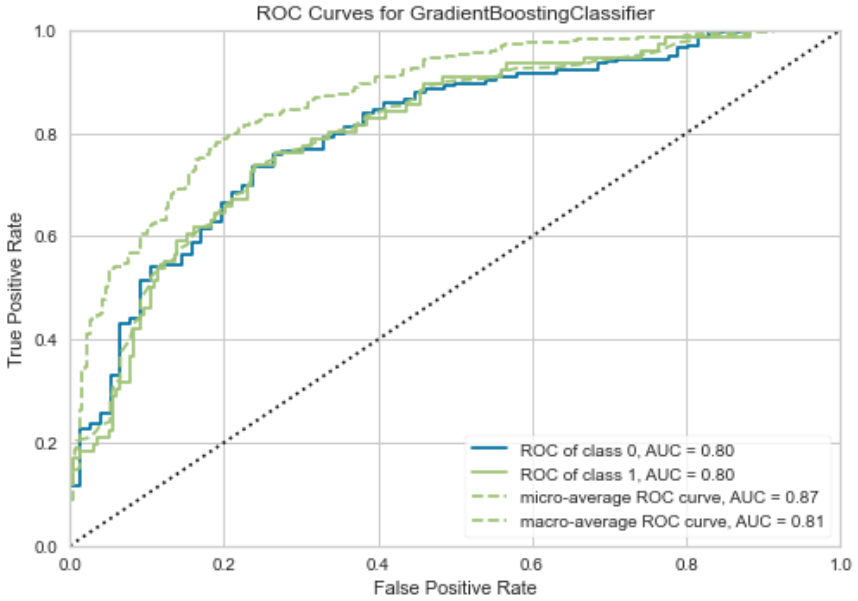
- Modeling was performed using python packages Sklearn, Tensorflow and Pycaret
- Our team decided to move forward with the Gradient Boosting classifier.

Model Comparison Summary

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
0	CatBoost Classifier	0.8265	0.8405	0.5608	0.6846	0.6062	0.4985	0.5086	1.9737
1	Gradient Boosting Classifier	0.8124	0.8401	0.6075	0.6255	0.6092	0.4873	0.4920	0.1602
2	Extreme Gradient Boosting	0.8096	0.8439	0.6186	0.6164	0.6122	0.4869	0.4906	0.1289
3	Light Gradient Boosting Machine	0.8039	0.8313	0.5382	0.6495	0.5705	0.4471	0.4613	0.3568
4	Extra Trees Classifier	0.7955	0.7902	0.4582	0.6182	0.5173	0.3937	0.4053	0.1441
5	Random Forest Classifier	0.7941	0.7772	0.4137	0.6267	0.4853	0.3677	0.3854	0.1123
6	Ada Boost Classifier	0.7785	0.8271	0.6118	0.5472	0.5732	0.4248	0.4299	0.0816
7	Logistic Regression	0.7418	0.8171	0.7657	0.4878	0.5930	0.4175	0.4443	0.0272
8	Decision Tree Classifier	0.7377	0.6557	0.4935	0.4848	0.4824	0.3087	0.3123	0.0077
9	Linear Discriminant Analysis	0.7320	0.8179	0.7546	0.4782	0.5826	0.4006	0.4261	0.0092
10	Ridge Classifier	0.7305	0.0000	0.7542	0.4756	0.5800	0.3971	0.4241	0.0075
11	SVM - Linear Kernel	0.7038	0.0000	0.7320	0.4412	0.5459	0.3459	0.3769	0.0095
12	K Neighbors Classifier	0.6318	0.6759	0.6699	0.3735	0.4760	0.2313	0.2554	0.0039
13	Quadratic Discriminant Analysis	0.3963	0.5982	0.9886	0.2904	0.4487	0.1055	0.2273	0.0068
14	Naive Bayes	0.3906	0.7524	0.9712	0.2864	0.4422	0.0952	0.2020	0.0042

Model Results

Gradient Boosting Classifier

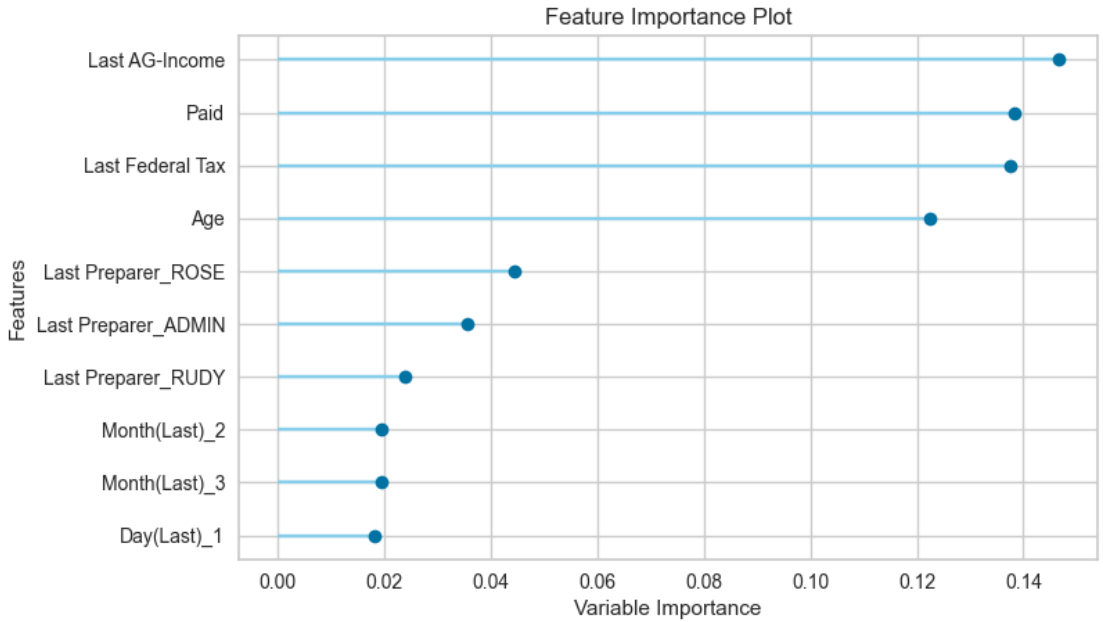


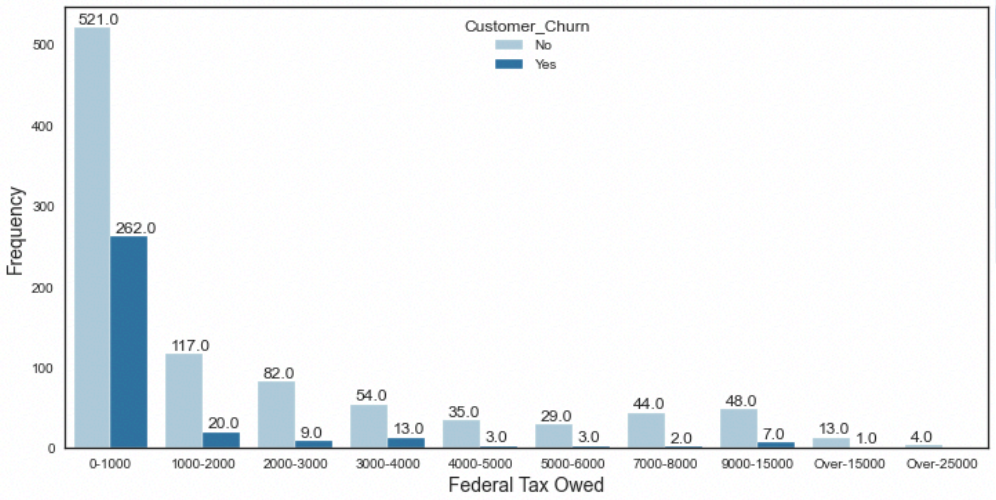
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Gradient Boosting Classifier	0.7934	0.8048	0.4868	0.6066	0.5401	0.409	0.4131

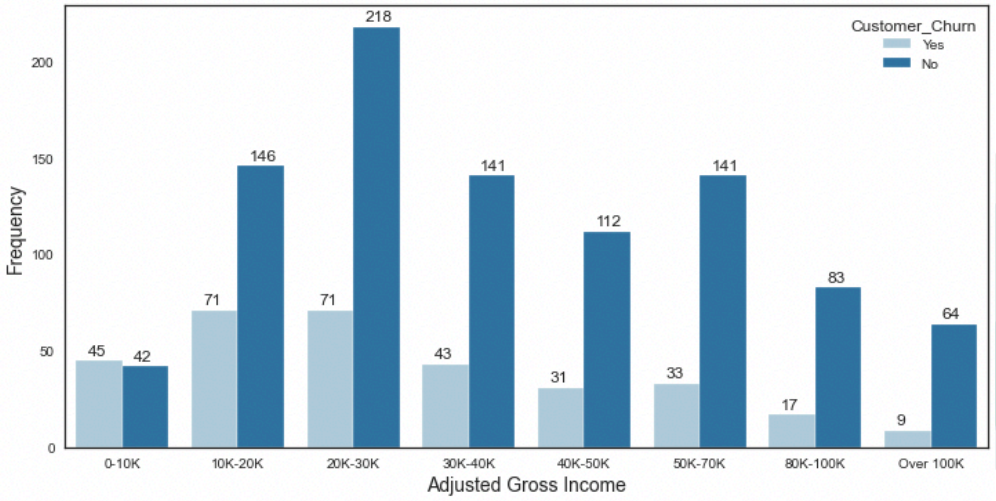
Analyzing Results

Selected Features for Analysis

- Adjusted Gross Income
- Fee Paid for Service
- Federal Tax Owed
- Customer Age
- Last Transaction (Day & Month)







Analyzing Results

Federal Tax Owed

- Most significant churn
  - $0-$2000 federal tax owed

Adjusted Gross Income

- Most significant churn
  - $10,000-$30,000 earned and taxable income

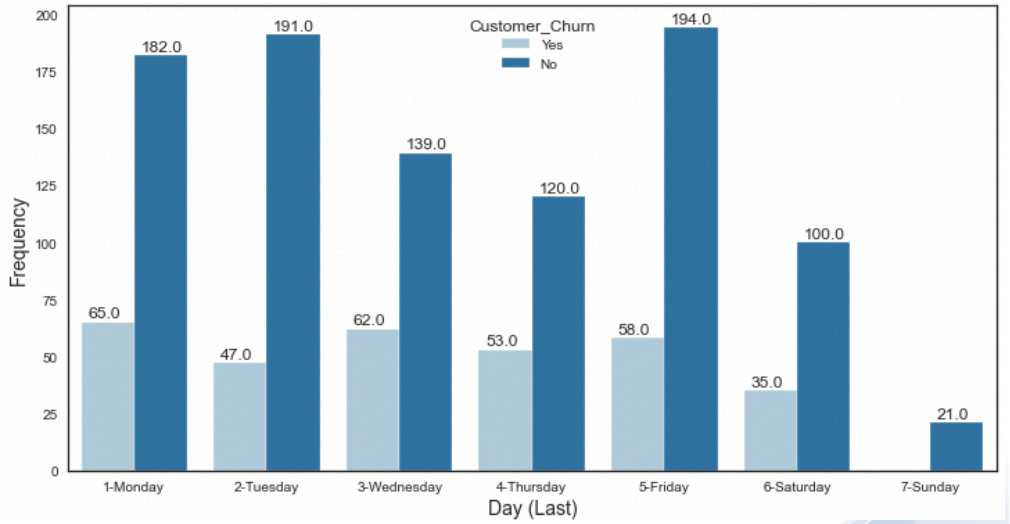
Analyzing Results

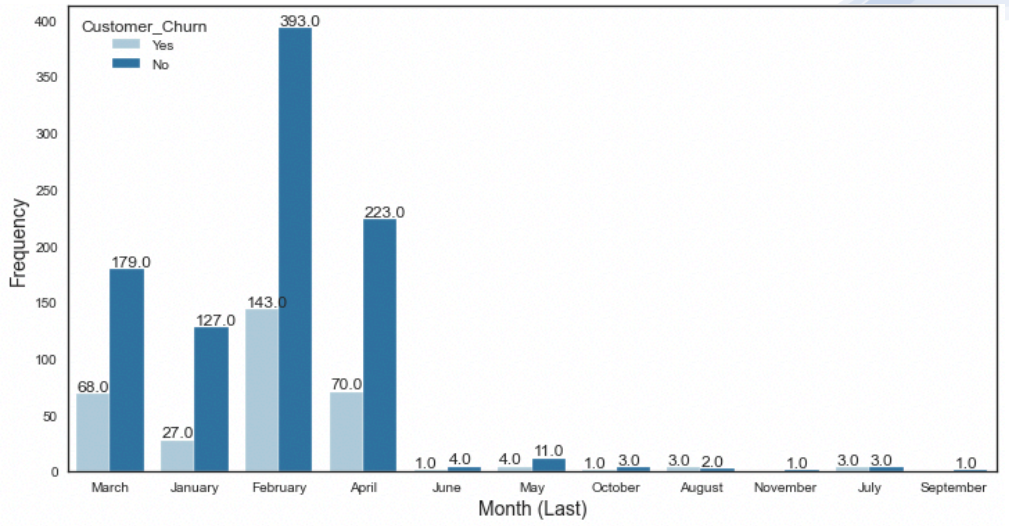
Last Day

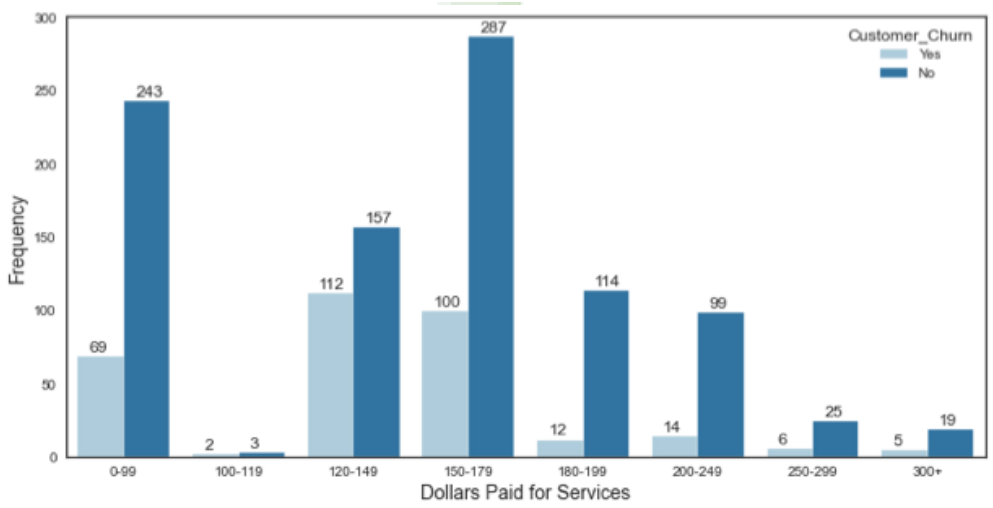
- Most significant churn
  - Wednesday and Thursday

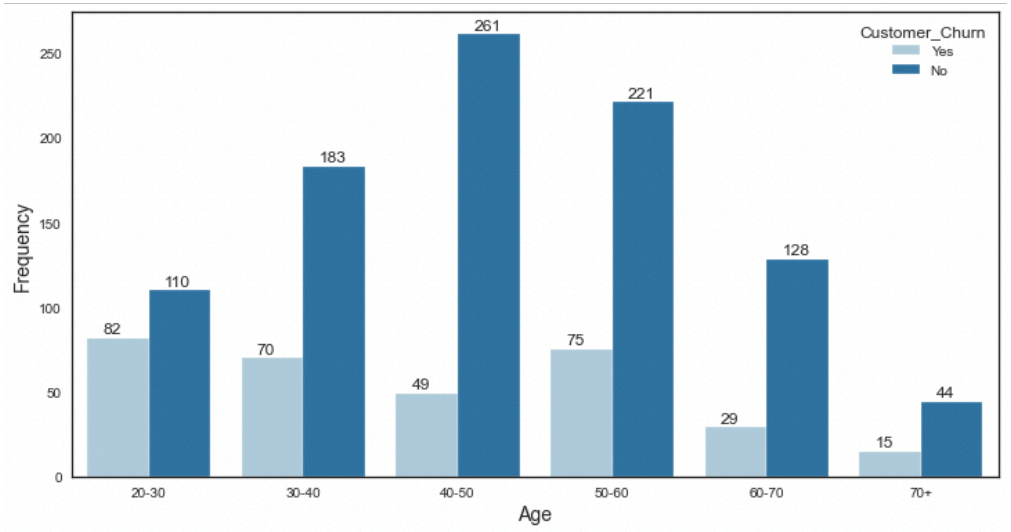
Last Month

- Most significant churn
  - March









Analyzing Results

Paid

- Most significant churn
  - $100-$150 fee for service

Age

- Most significant churn
  - Customers ages 20-30

Business Recommendations

- **Target Market -** The tax office needs to target younger consumers, specifically those that are between 20-30 years old and live in single income households. To help build rapport with the younger demographic, the Tax office should begin offering financial wellness checks to those that are interested. A simple check of finances, plans for retirement and other provided resources can go along way when looking to build trust with a client base.
- **Technology** - In addition, the office should invest in video conferencing software such as Zoom, GoToMeeting or Google Meet in order to offer alternative methods of tax preparation to their client base and offer a user friendly approach to their younger demographic.
- **Update Fees -** Reclassify listed fee amounts. Prices actually paid tend to be lower than those that are listed. The price will appear as a reduction in cost but the variation in fees collected will be minimal.
- **Influence Traffic -** Redistribute the large amount of client traffic that occurs in February through outreach initiatives, online servicing or price discounts. Leverage Sundays for several of the tasks above (low traffic, no customer churn).

Future Work | Limitations

- Future Work
  - Comprehensive data per client
    - Number of jobs per client
    - Additional breakdown of income earned
      - Investment earnings
      - Taxable wages
    - Healthcare Status
      - Insured vs. not insured
    - Education level

- Limitations
  - Poor record keeping resulted in loss of potentially important features
  - 2019 Tax year data not included

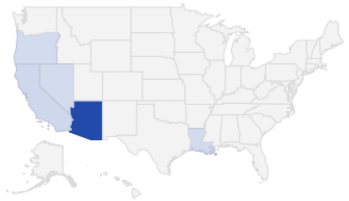
THANK YOU!

Questions? Miguel Santana | contact: msantana269@gmail.com

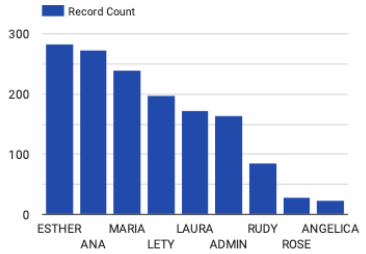
Appendix

Tax Office Executive Dashboard (Overview)

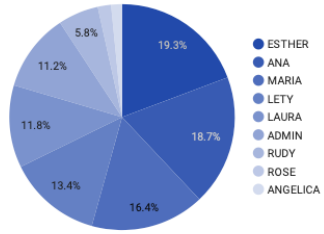
Total Clients By Region (2016-2018)



Total Clients By Preparer (2016-2018)



Percentage of Traffic



	City	Count ▾	%
1.	YUMA	1,096	74.71%
2.	SOMERTON	190	12.95%
3.	SAN LUIS	91	6.2%
4.	WINTERHAVEN	17	1.16%
5.	WELLTON	9	0.61%
6.	TUCSON	5	0.34%
7.	SAN DIEGO	4	0.27%
8.	SALINAS	4	0.27%

1 - 43 / 43

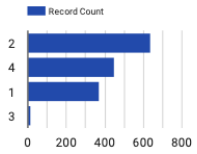
	Last Preparer	Count ▾	%
1.	ESTHER	283	19.29%
2.	ANA	274	18.68%
3.	MARIA	240	16.36%
4.	LETY	197	13.43%
5.	LAURA	173	11.79%
6.	ADMIN	164	11.18%
7.	RUDY	85	5.79%
8.	ROSE	28	1.91%

1 - 9 / 9

E-File Scorecards

2016 E-File Count: 993	2016 Paid: 115,058
2017 E-File Count: 1,238	2017 Paid: 154,627
2018 E-File Count: 1,147	2018 Paid: 151,865

Clients By Family Size



	Family Size	Count
1.	1	25.15%
2.	4	30.54%
3.	2	43.42%
4.	3	0.89%

1 - 4 / 4

Clients By Adjusted Gross Income

	Adjusted Gross Income	Record Count	%
1.	50,000-69,999	192	13.09%
2.	20,000-29,999	332	22.63%
3.	10,000-19,999	259	17.66%
4.	0-9,999	110	7.5%
5.	40,000-49,999	164	11.18%

1 - 8 / 8

The tax office executive dashboard is available via google data studio.

You can find it here: https://datastudio.google.com/reporting/20efa577-3f3f-448c-a11b-0ccbba069500/page/U6UqB