Data Driven Solutions:

A practical overview on Machine Learning

Startup Weekend
Raul Eulogio

Introduction

- Raul Eulogio
  - Data Analyst at **Hospice of Santa Barbara**
  - Co-founder: inertia7.com
  - President of Data Science at UCSB
  - Self taught **Machine learning** enthusiast

Data Science at UCSB and Farmer's Insurance Competition

Farmers Insurance is challenging you to put your data skills to the test. Seize this opportunity to practically apply data science to tackle a problem in the insurance field.

The top-performing teams will bring home:

- 1st place: $2000
- 2nd place: $1000
- 3rd place: $500

Additionally, all winning teams will get to present their work to a panel of Farmers employees. **MUST BE UCSB Student and Paid Member of Data Science at UCSB**. Application here

Why Machine Learning*?

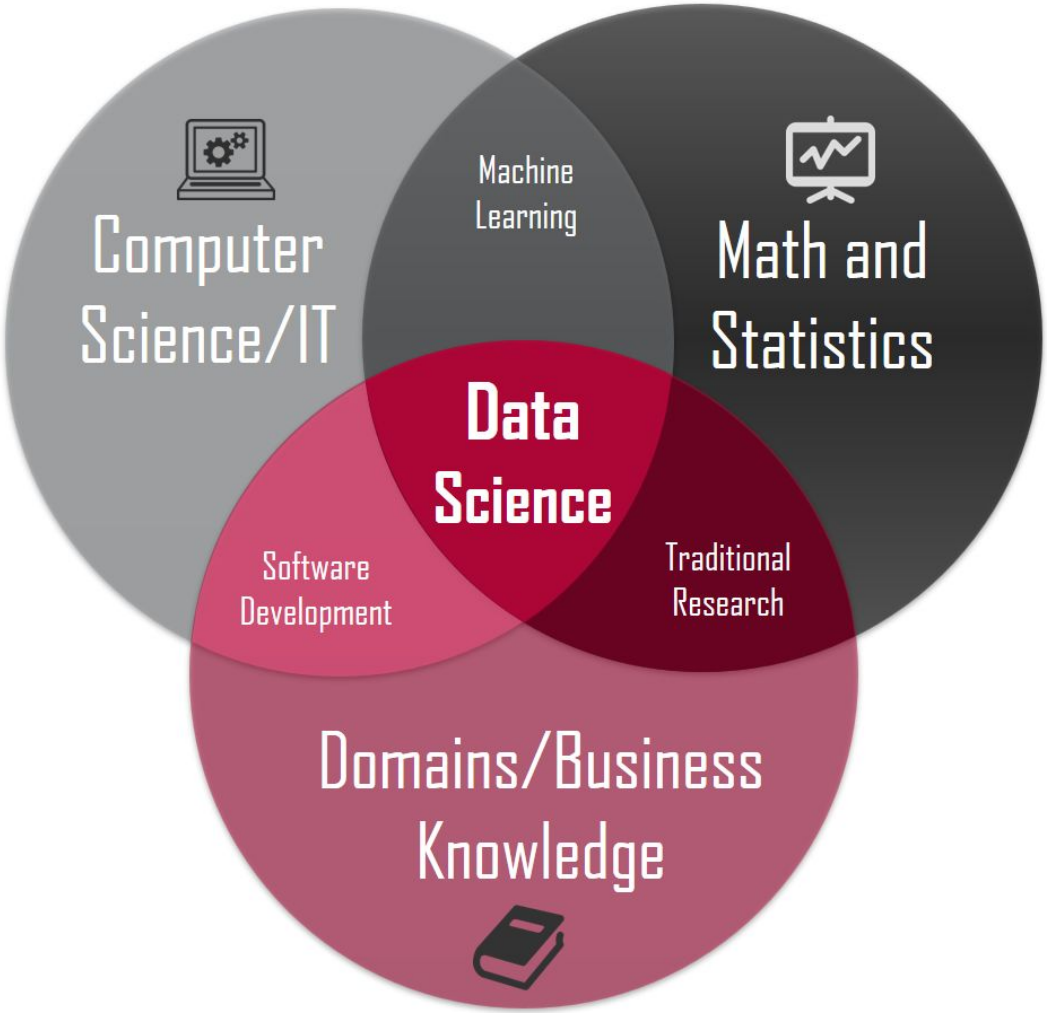
- Use data collection to your advantage
  - The authors of Lean Analytics state: **“*data driven learning is the cornerstone of success in startups. It’s how you learn what’s working and iterate towards the right product and market before the money runs out.*”**
- Data Science
  - Enhancing the interpretation of reality
  - Automating machines to respond to their environments

* I will use Machine Learning and Data Science interchangeably

Data Driven Solutions for a **Data Driven Organization** in a **Data Driven World**

I’m not here ...

- to tell you **Data Science** *“is the sexiest job of the 21st century”*
- to tell you that studies show *less than 1% of data is being analyzed*
- to show you the usual **Venn Diagram** that is presented at almost every data science talk


Computer Science/IT
Machine Learning
Math and Statistics
Data Science
Software Development
Traditional Research
Domains/Business Knowledge

Multifaceted Domain

Machine Learning can be ...

- Exploratory Analysis
  - Exploring trends in data
  - Creating a narrative with data
- Unsupervised Learning
  - Exploring Trends and Patterns on a larger scale
  - Find hidden structure in data
- Supervised Learning
  - Predicting an output based on inputs
  - Regression and Classification

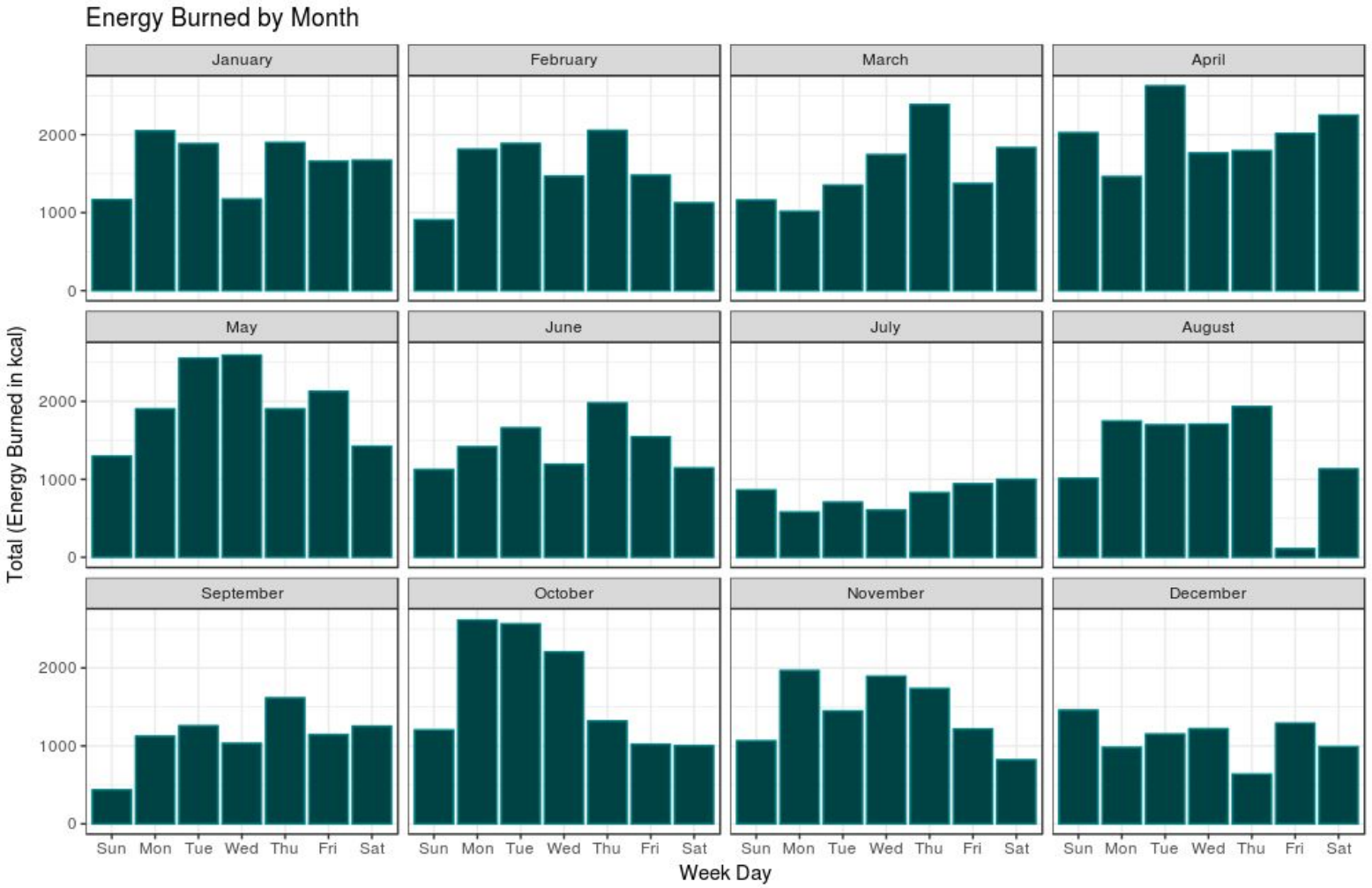
Case Studies

Show through examples, all data available online and all work is open source and on my Github Repository!

- Exploratory Analysis - Apple Watch Data
  - Existing data collected by customer/user
    - Data Collection (**Python**) and Data Exploration (**R**)
- Unsupervised Learning - Spotify Data
  - Data made available by 3rd party source
    - Data Exploration and Data Modeling (**Python**)
- Supervised Learning - IBM Customer Churn Data
  - Data collected by organization
    - Data Exploration (**R**) and Data Modeling (**Python**)

Exploratory Analysis: A case study on Apple Watch

- Sisense states: "***You do [EDA] by taking a broad look at patterns, trends, outliers, unexpected results and so on in your existing data, using visual and quantitative methods to get a sense of the story this tells. You're looking for clues that suggest your logical next steps, questions or areas of research.***"
- Example data was gathered by *Apple Watch* on a daily basis to help fitness tracking and other health related data


Energy Burned by Month
January
February
March
April
May
June
July
August
September
October
November
December
Total (Energy Burned in kcal)
0
1000
2000
Sun
Mon
Tue
Wed
Thu
Fri
Sat
Week Day

Questions to consider

How can we use customer's daily fitness regimen to identify important trends?

Can we detect and prevent days/weeks where our customers will reduce their workout regimen?

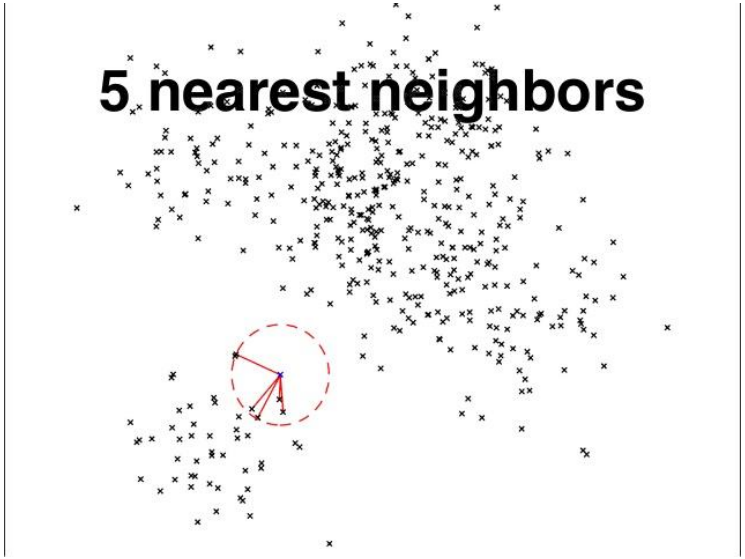
Is there correlation between our customer's workout regimen and use of our services?

Unsupervised Learning: A Case Study on Spotify Music

- Noticing trends and patterns within the data
  - Combining all features and usually unlabelled data
- Using Spotify API to create a recommender based on distance metric
- Ability to create clusters within our data
  - Recommend other use cases/product based on customer preference
    - Examples include *Recommended videos* on Youtube, *Customers also bought* on Amazon, *Daily mix* on Spotify

How does it work?

- Nearest Neighbor Algorithm
- Algorithm creates feature space using all inputs
- Inputs include:
  - Danceability
  - Loudness
  - Tempo
  - Category
  - And more



Examples of recommender at work

```
In [14]: return_recommendation(hi_lo)
```

```
Recommendations for Ooh La La by HI-LO
0. Recommended Song: More Mess - Hugel Remix by Kungs
1. Recommended Song: Animals - Victor Niglio & Martin Garrix Festival Trap Mix by Martin Garrix
2. Recommended Song: Brolab by Tiësto
3. Recommended Song: Real Love - Radio Mix by Antonio Giacca
4. Recommended Song: Revolt by Tiësto
5. Recommended Song: Boombox by Dirtyphonics
6. Recommended Song: Make You Hustle by Croatia Squad
7. Recommended Song: Imjussayin by Convex
8. Recommended Song: Get Down by Hardwell
```

After some research I found that a lot of these songs were very similar in nature!

```
Recommendations for All We Got (feat. Kanye West & Chicago Childrens Choir) by Chance The Rapper
0. Recommended Song: Perplexing Pegasus by Rae Sremmurd
1. Recommended Song: Runaway Train by Soul Asylum
2. Recommended Song: Too Hotty by Various Artists
3. Recommended Song: Back (feat. Lil Yachty) by Lil Pump
4. Recommended Song: Summertime Sadness by Lana Del Rey
5. Recommended Song: New Freezer (feat. Kendrick Lamar) by Rich The Kid
6. Recommended Song: Dig Down by Muse
7. Recommended Song: Dont Wanna Know - Acoustic Version by The Mayries
8. Recommended Song: Follow You - Tep No Edit by Jamie Brown
```

I have more knowledge in hip hop so I can say many of these were good recommendations some interesting songs were **Don’t Wanna Know** and **Summertime Sadness.**

Questions to consider

How can we provide a seamless music experience for users?

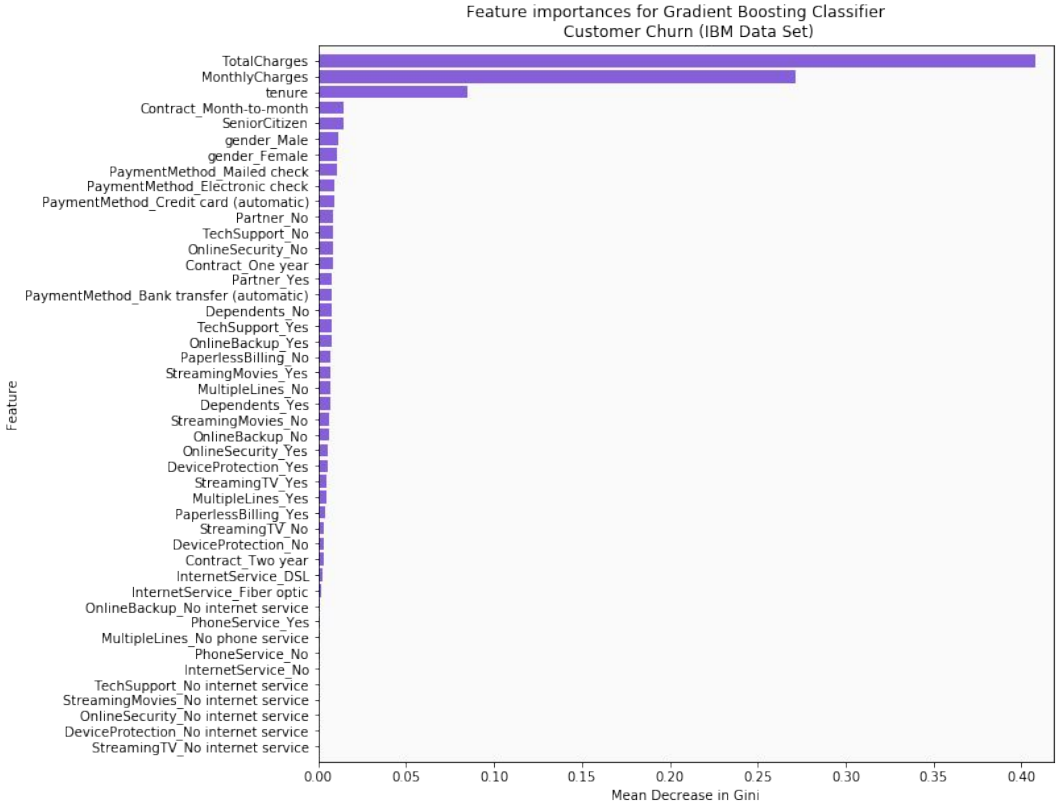
Can we understand a users musical taste to maximize daily workout regime?

How can we effectively utilize 3rd party data to benefit our product?

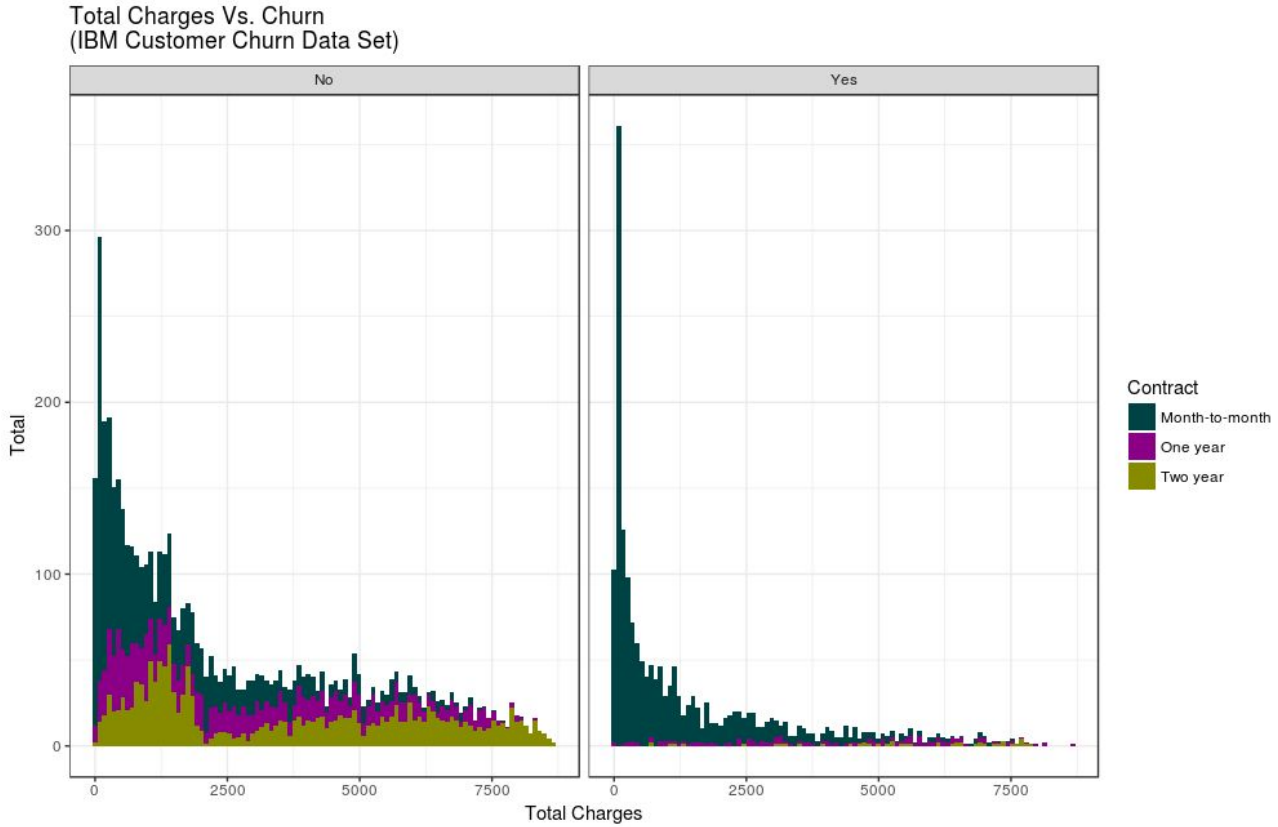
Supervised Learning: A case study on Churn Rate

- Can we accurately predict when a customer will stop using a service/business?
- Binary Classification problem using customer information including:
  - Tenure
  - Total Charges
  - Monthly Charges
  - Gender
  - Utilization of Phone Services?
  - And more...
- Models used:
  - Gradient Boosting
  - Logistic Regression
- Things to consider: Data Preprocessing, Data leakage, Class Imbalance

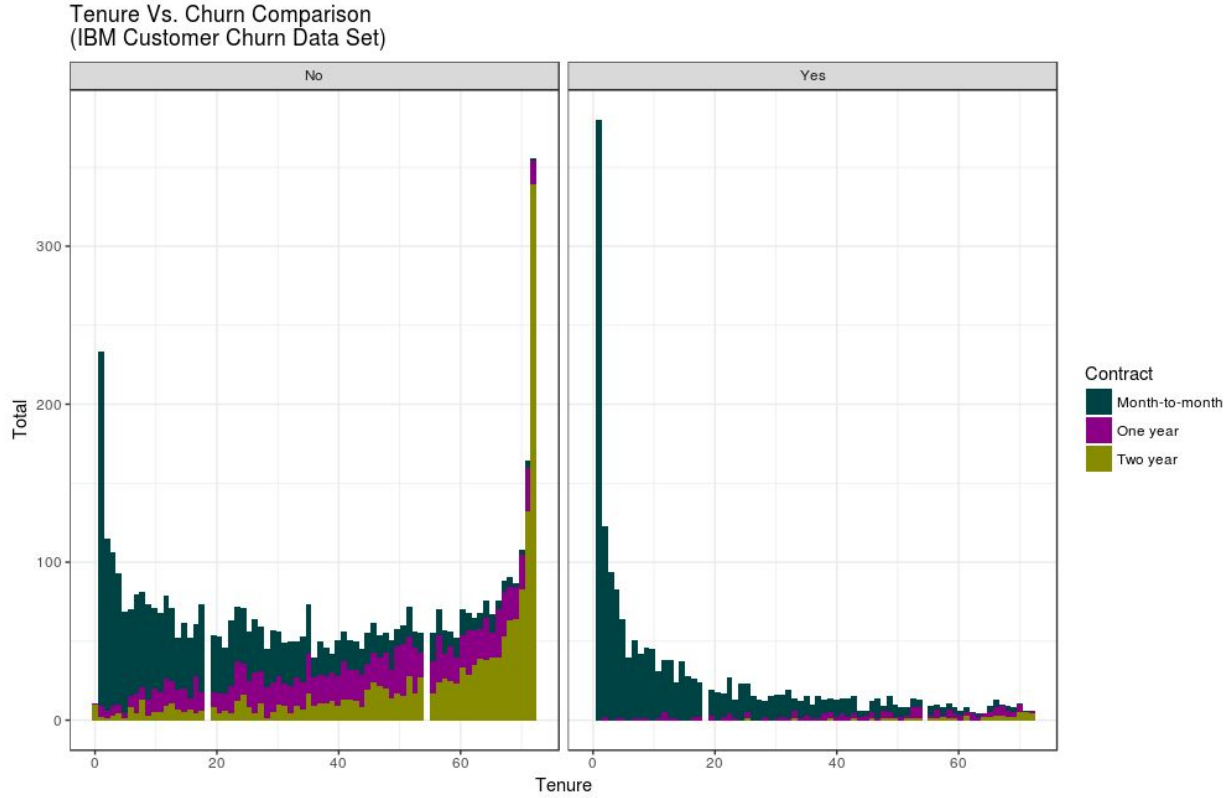
Variable Importance gathered by Gradient Boosting



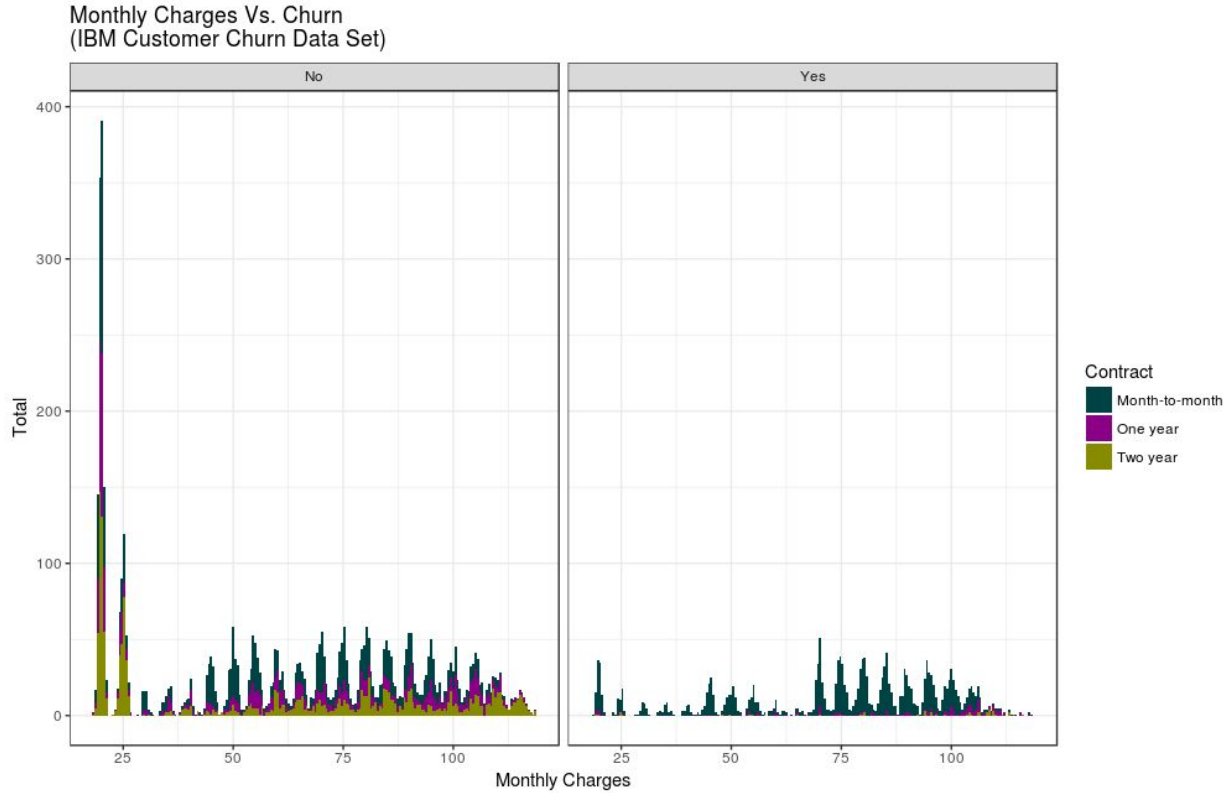
Exploratory Analysis on Churn Data Set



Exploratory Analysis on Churn Data Set

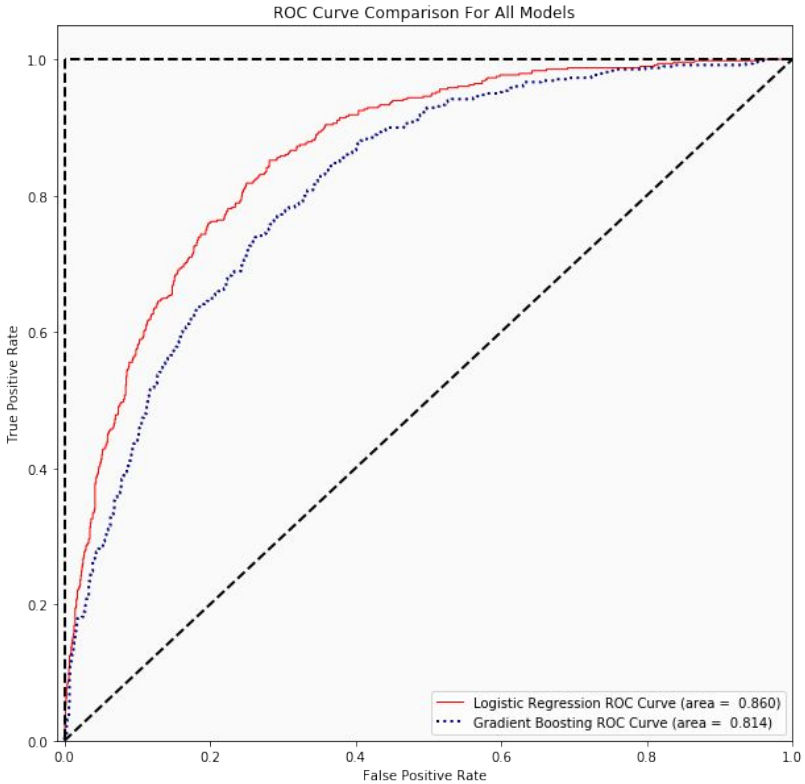


Exploratory Analysis on Churn Data Set



Final Results on Churn Data Set

- Gradient Boosting: 76% Accuracy (**CV**)
- Logistic Regression: 77% Accuracy (**CV**)
- Not high but we can still gain insight
  - Variable Importance for GB
  - Coefficients for variables for LR
- Customers with Month-to-Month Contracts most likely to *Churn*
- Neural Networks? Careful of Black Box Model
- Collect more data!



Results

- “*All models are wrong but some are useful*" - George Box
- ~77% accuracy for both *Logistic Regression* and *Gradient Boosting*: Not too high in terms of groundbreaking results but can still can give insight. Typically 90% accuracy is a good start
- Key Takeaway: Data Science/Machine Learning is **a life cycle not a one and done procedure**.
- Iterations are key; if model and data didn’t output wanted results, collect more data. Ask what data should be collected and how it should be collected with key stakeholders.

Questions to consider

How can we integrate Customer Reviews into our Machine Learning process?

What other covariates can we consider when creating our models?

Are we collecting the right data?

Which model can give us the most insight into our data without being to computationally expensive?

Q&A

- If you have any questions or would like to contribute to these projects email me: raul.eulogio@inertia7.com
- Check out inertia7.com if you want to learn all things Machine Learning and Data Science

Resources

- Overview of Machine Learning using scikit-learn
- Introduction to Gradient Boosting
- Book Recommender (Inspired Spotify Recommender)
- Github Repo with Source code for presentation
- Logistic Regression with Scikit-learn