**Presentation**

1. Names and ID

Yosef Gerstein 327212163

David King

Liora Kala

1. The Problem
   * **What are you trying to predict?** 
     1. The predicable characteristics of the customers that have the highest churn rate.

* **What data is available for that (input)?**
  + Churn and Demographic (gender, senior citizen, partner, dependent)
  + Churn and product engagement-product performance vis-à-vis churn (phoneservice , internetservice, onlinesecurity, onlinebackup, deviceprotection, techsupport, streamingtv, streamingmovies)
  + Churn and seniority (tenure, contract)
  + Churn and expenditure (monthlycharges, totalcharges, paymentmethod, paperlessbilling)
* **What can be the motivation and applications for solving the problem?**
  + Motivation
    - Identifying the problem’s importance – To define which primary issue(s) are causing the most churn to focus organizational priority in preventing churn.
    - Urgency and Impact – The organization can significantly prevent loss, increase customer experience, and ultimately customer revenue through positive experience that increases engagement to product.
    - Innovation and Advancement – Discover organizational and product deficiencies that could lead to organizational restructuring for efficiency, product development priorities, market opportunities, and new customer engagement protocols to increase customer value.
* Applications
  + Real-World Use Cases – The trending product which customers are most engaged to
  + Industry Relevance – This is applicable to SAS and broadly to any service and payment processing industries.
  + **Cross-disciplinary Applications – Customer and organizational behavioral predictive analytics**

1. **Data Description**
   * **How many examples in the dataset?** 
     1. We were measuring relationship with churn and demographics, product engagement/performance, support engagement, seniority in company, expenditure. 7043 rows
   * **How many features are in the dataset?** 
     1. 21 columns, that we were able to reduce to 17 when combining columns(streaming services), and removing customerID
   * **What is the distribution of the labels?**
     1. 5174 retained and 1869 churned.
   * **Are there any missing values?**
     1. Yes, there are missing values in totalcharges (11)
   * **Show 5 tp 6 graphs describing various aspects of the data**
     1. 20 graphs!
2. **Data Engineering** 
   * **Did you remove any features?**
     1. Removed only CustomerID and paperless billing does not matter, as we will have to provide a bill regardless and based on the data the organization offers both choices, as this was not relevant, every other feature was used to measure.
   * **Did you add any features?**
     1. While we did not any unique features, we combined columns that seem to logically connect, like the various steaming services (eg. Steaming movies and Steaming Tv was combined into column steaming)
   * **What did you do with missing values?**
     1. I converted the missing values from string to NaN, and because the data set is relatively small and the NaN was small (11), I enumerated and estimated the values based upon the corresponding values.
3. **ML Algorithms: test the following algorithms:**
   * K nearest neighbors, try 3 different k values
     1. Created 4. Found that 9 neighbors created the highest reliability (0.77)
   * Decision tree with 3 different max depth values
     1. Created 5. Found that the depth of 6 gave us the most reliability (0.78).
   * Random forest with 3 different max depth values and 100 estimators
4. **Algorithms introspection – inspect the various algorithms artifacts.**
   * What decision trees did you get?
     1. With our original depth of 3, we received a decision tree of 1 primary node, 6 child nodes (branches), and 8 leaves
   * What is the random forest feature importance?
     1. The random forest feature is important in that it provides a list of the most important independent variables ranked from the highest correlation to our dependent variable of Churn