# Project One Starting page:

Hello! My name is Sydney Bailey and I am very excited to be with you today to discuss our findings. I am a data scientist with Data Analytic Solutions. I will be presenting today on understanding key drivers that contribute to employee turnover.

# Agenda

Here is a look at our agenda today. We will be going through an introduction of why we are interested in employee attrition, then look at the data analysis and highlighting key insights contributing to attrition at Frito Lay, then we will look at a couple of the models that were utilized to use those key insights and predict turnover, then lastly we will explore the cost implications of what we can do to mitigate the attrition.

We’ll start with an introduction to why employee attrition is an important focus. Next, we’ll dive into the data analysis and highlight the key factors driving attrition at Frito-Lay. Then, we’ll review a few predictive models that leverage these insights to forecast turnover. Finally, we’ll explore the potential cost implications and strategies for mitigating attrition.

# Introduction – Understanding Employee attrition through data

Lets start with a reminder of why we are here today. As I stated earlier, Frito Lay partnered with Data Analytic Solutions to figure out what is driving employee turn over at the company. The company has collected information for the last year and provided it to Data Analytic Solutions to analyze. Our objective is to provide the key indicators of attrition, a way to predict at risk employees, and a financial analysis.

As we move to the next slides, we will start to discuss our analysis of the variables impacting attrition starting with the most impactful.

# Job Involvement Strongly Affects Attrition

I am going to go into a bit more detail on this slide around the analysis that was performed for each variable. First, I looked at the graph of the attrition rate. As you can see here, attrition rate was extremely high for low job involvement. It was also double the next highest level. This was a key that made me wonder if I were to run a statistical test around significance how much it would be. I chose to run a Chi-Squared test on the data. The chi-squared test is a statistical hypothesis test for analyzing categorical data. It determines if there is a statistically significant association between two categorical variables in a contingency table. With this test, the higher the Chi-Squared value means there is a large difference between the observed and expected frequencies. This suggests that the difference is not likely due to random chance. The lower the p-value (typically ≤ 0.05), You conclude that there is a statistically significant relationship or difference between the categories. Since there is a high Chi-Squared value, and a very low p-value we conclude that there is a statistically significant relationship between attrition and job involvement. This means that the lower the job involvement, the higher the likelihood of leaving the company.

# Attrition Significantly Differs by Marital Status

Next we are going to look at Marital Status. As I graphed marital status by attrition percentage, I noticed the significant amount of attrition for the employees who were single. This was also a very high chi-squared value meaning that it was not due to random chance, especially when you see how low under 0.05 the p-value is. This implies that there is a higher likelihood of people leaving who are single.

# Lower Income Associated with Higher Attrition

Lastly we are going to deep dive monthly income. If you look at the box plots, you can see that the line in the box plot (or the median) is much lower for people who left the company. Meaning that people who left earned less than people who stayed. When I ran a test on the values, it showed that t value is quite high and the p value is very low. This implies that there is a relationship between pay and retention. In fact, employees who left earned less per month than those who stayed.

# And Many More

These were not the only ones that were analyzed. Many more variables were analyzed to determine the key drivers of turn over.

Spotting the Key Predictors of Attrition

In fact, we did an analysis of each variables and put them together from lowest P value and highest t or chi squared values. By doing this, we were able to determine that the most impactful drivers to turn over were Job Involvement, Marital Status, and Monthly Income. We used a set of variables in our analysis as noted. So lets dive into our analysis.

# Evaluation of the K Nearest Neighbor Model K = 5

We took the initial set of data and ran the K Nearest Neighbor model. As you can see, we predicted quite a lot of people to not leave and most employees to stay. In fact, our sensitivity was only 5.77% and Specificity at 96.65%. But what does this mean for attrition predictions. It means that this model will predict 5.77% of people who will leave the company. The good news is that we will predict almost everyone who will stay. This is not an ideal model. So lets look at another modeling type.

# Evaluation of the Naïve Bayes Model

As you can see here, this model did slightly better at predicting people who will leave. Our sensitivity came in at 25 percent and specificity at 95%. So this model predicts 25% of people who will leave and almost everyone who will stay. But… based on the ask of the customer this is not a successful model. We really want to increase our sensitivity so that we can predict those who will leave and thus allowing us to mitigate those. So what actions can we take to help increase our sensitivity.

# Synthetic Minority Over-sampling Technique Overview

* SMOTE is a data preprocessing technique used in machine learning to address class imbalance.
* It creates synthetic examples for the minority class to equalize the class distribution, helping to prevent models from becoming biased towards the majority class.
* We used this technique for this data set because we had 730 data points for people who did not leave the company and 140 data points for people who did leave the company.

# Evaluation of the K Nearest Neighbor Model K = 5 (smote)

After re-sampling the data we have a huge improvement in our KNN model. We have increased our sensitivity and our specificity! This means that we will successfully predict 60% of people who are at risk of leaving the company while also predicting 59% of people who will stay! This is a huge increase from our original model around the most important variable to more accurately predict who is at risk of leaving. So lets see how our Naïve Bayes model did with re-sampling since it was already performing better without re-sampling.

# Evaluation of the Naïve Bayes Model (smote)

This is also a huge improvement from our original model by increasing both sensitivity and specificity. We are predicting almost 70% of people who are at risk of leaving accurately while keeping a very high 62% of those who will stay which is a huge increase from the 25%. This model is the clear winner between the two models we have based on the number for sensitivity. Now that we have improved both of our models, lets look at the cost and savings portion of our analysis.

# Assumptions for Cost of leaving calculation

Based on the information provided for the cost of employee turnover, we are going to be using 9 months of an employees salary if we were to have to replace someone who has left the company. We took the average salary across all job levels for the cost calculation. Next we look at the employee incentives to look at how we mitigate employees from leaving. We are going to offer employees who are at risk of leaving $10,000 to stay. If they accept this offer, they are required to stay an additional 2 years or they will be required to pay back the money they accepted. This will reduce the number of people who leave by 31% on average.

# Cost Evaluation of the KNN Model

As we remember from our model, our sensitivity is sitting at 60% for predicting those who are at risk of leaving and our specificity is sitting at 62% for predicting those who will not leave. When we calculate our F1 score, it shows that we are predicting quite a few false positives which you can see at the 83. Lets see how this impacts cost. So the cost associated with no interventions is the 41 people who leave times the 225 thousand. So that comes to 9.2 million dollars to replace the people who leave. So, if we use our mitigation of long term incentives to the 108 people we predict to leave, and then add in the 16 who leave who we predicted wouldn’t plus the 25 who will leave \* 31% reduction which equals 17 multiplied by the 225 you get 8.5 million. So the total cost savings is 9.2 million minus 8.5 million which comes to 720K savings. Not bad! Almost three quarters of million dollars saved by using long term incentives. Lets look at our Naïve bayes model which performed better than the KNN model.

# Cost Evaluation of the NB Model

As we remember from our model, our sensitivity is sitting at 70% for predicting those who are at risk of leaving and our specificity is sitting at 65% for predicting those who will not leave. When we calculate our F1 score, it shows that we are predicting quite a few false positives which you can see at the 75 which is less than the KNN model. Lets see how this impacts cost. So the cost associated with no interventions is the 47 people who leave times the 225 thousand. So that comes to 10.6 million dollars to replace the people who leave. So, if we use our mitigation of long term incentives to the 109 people we predict to leave, and then add in the 13 who leave who we predicted wouldn’t plus the 34 who will leave \* 31% reduction which equals 23 multiplied by the 225 you get 9.2 million. So the total cost savings is 10.6 million minus 9.2 million which comes to 1.4 million in savings. Huge improvement from the KNN model! 1.4 million dollars saved by using long term incentives.

# From Data to Action: Retaining Our Talent

So what did we learn from our analysis! We learned there were several key reasons that people were leaving. The most impactful was job involvement, followed by marital status, and then monthly income. We are seeing that younger, less experienced employees were more likely to leave. If we were to start offering incentives to these employees at high risk of leaving based on our prediction, we could save the company around 1.4 million dollars. Overall we believe its most important to focus on early career employees and building the relationships and roots to enable them to feel a part of something bigger than themselves.