**PROJECT PROPOSAL**

**1. What is the problem you want to solve?**

“I QUIT!”, is the last thing a company wants to hear from their employees. In a sense, it’s the employees who make the company. It’s the employees who do the work. It’s the employees who shape the company’s culture. Long-term success, a healthy work environment, and high employee retention are all signs of a successful company. But when a company experiences a high rate of employee turnover, then something is going wrong. This can lead the company to huge monetary losses by these innovative and valuable employees. My goal is to understand what factors contribute most to employee turnover and create a model that can predict if a certain employee will leave the company or not.

**2. Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn’t have otherwise?**

Companies that maintain a healthy organization and culture are always a good sign of future prosperity. Recognizing and understanding what factors that were associated with employee turnover will allow companies and individuals to limit this from happening and may even increase employee productivity and growth. These predictive insights give managers the opportunity to take corrective steps to build and preserve their successful business.

**3. What data are you going to use for this? How will you acquire this data?**

The data was found from the “Human Resources Analytics” dataset provided by Kaggle’s website. <https://www.kaggle.com/ludobenistant/hr-analytics>

**4. In brief, outline your approach to solving this problem (knowing that this might change later).**

I’ll be following a typical data science pipeline, which is “OSEMN” (pronounced awesome).

Obtaining the required data is the first approach in solving the problem. I would have to download the dataset from Kaggle’s website and import it as a csv file to my working environment.

Scrubbing or cleaning the data is the next step. This includes data imputation of missing or invalid data and fixing column names.

Exploring the data with exploratory data analysis will follow right after and allow further insight of what our dataset contains. Looking for any outliers or weird data. Understanding the relationship each explanatory variable has with the response variable resides here and we can do this with a correlation matrix. The creation or removing of features through the use of feature engineering is a possibility. The use of various graphs plays a significant role here as well because it will give us a visual representation of how the variables interact with one another. We will get to see whether some variables have a linear or non-linear relationship. Taking the time to examine and understand our dataset will then give us the suggestions on what type of predictive model to use.

Modeling the data will give us our predictive power on whether an employee will leave. Types of models to use could be RF, SVM, LM, GBM, etc. Cross validation is used here, which will allow us to examine our model’s accuracy and tune our model’s hyperparameters if necessary. We can also use some feature selection from RandomForest. A confusion matrix can give us our precision of our model with the number of True Positives and True Negatives. We can graph this with a ROC curve. Understand the reasoning behind choosing the right model for this problem.

Interpreting the data is last. With all the results and analysis of the data, what conclusion is made? What factors contributed most to employee turnover? What relationship of variables were found? If our model’s accuracy is too high from our test set, a chance of overfitting is likely. Ways to prevent overfitting include: collecting more data, choosing simpler models, cross validation, regularization, use of ensemble methods, or better parameter tuning. Give a brief overview of the feature importance that affected our model. How can we improve our model in the future?

**5. What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.**

My deliverables would be my python notebook, which includes the source code, visualizations, and some documentation.