

Doing Data Science: Case Study 2

Frito Lay – Employee Attrition & Salary Prediction

Sterling Beason, Data Scientist @DDSAanalytics
2019-12-05

Objective

Analyze & build accurate predictive models for employee attrition and salary (monthly income) for Frito Lay.

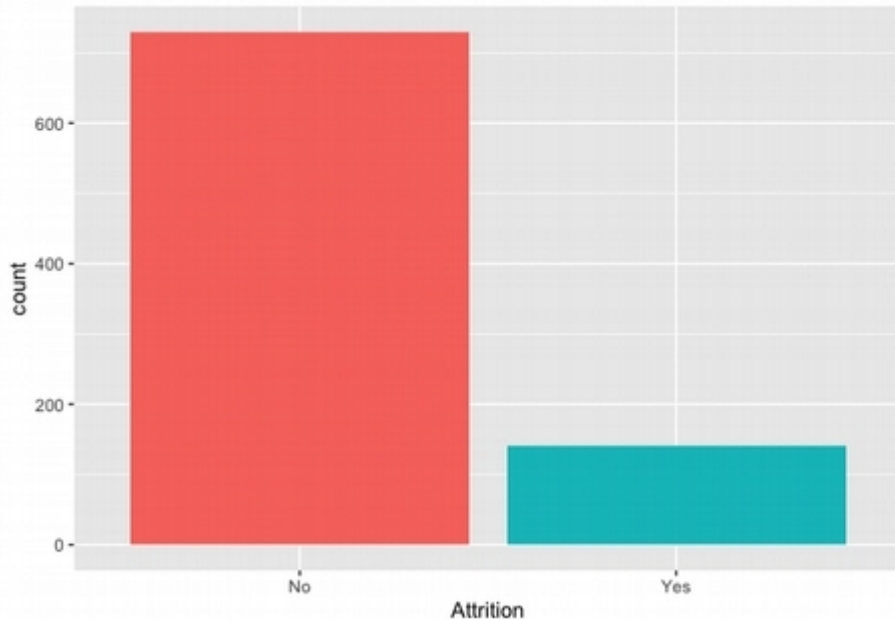
Dataset

For our analysis and modeling, Frito Lay provided the ***CaseStudy2-data.csv*** dataset. This dataset includes features on employees; such as, job levels, stock option levels, total working years, overtime required, etc...

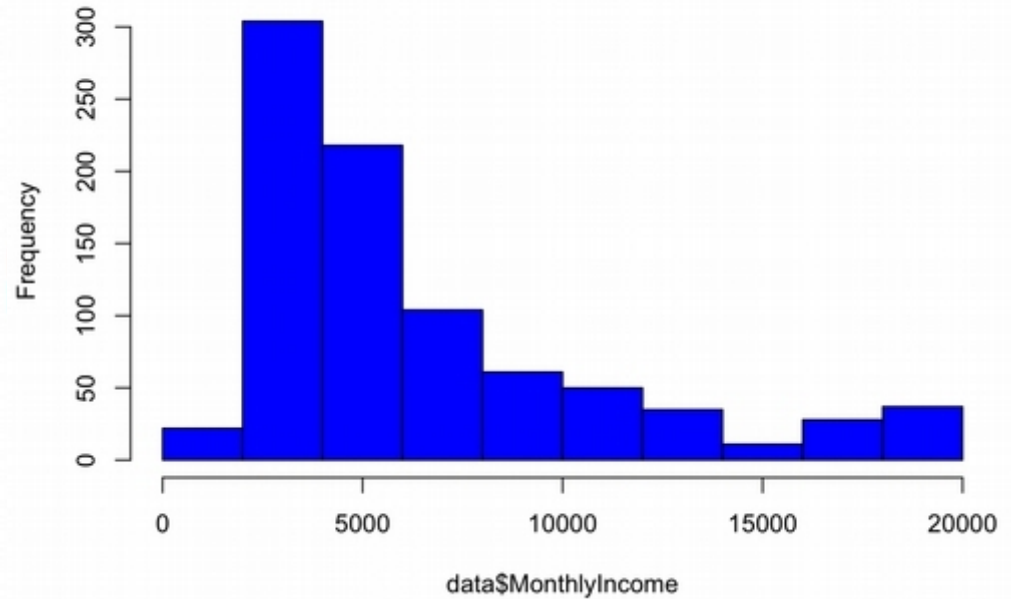
- 870 observations
- 36 features

Exploratory Data Analysis (EDA)

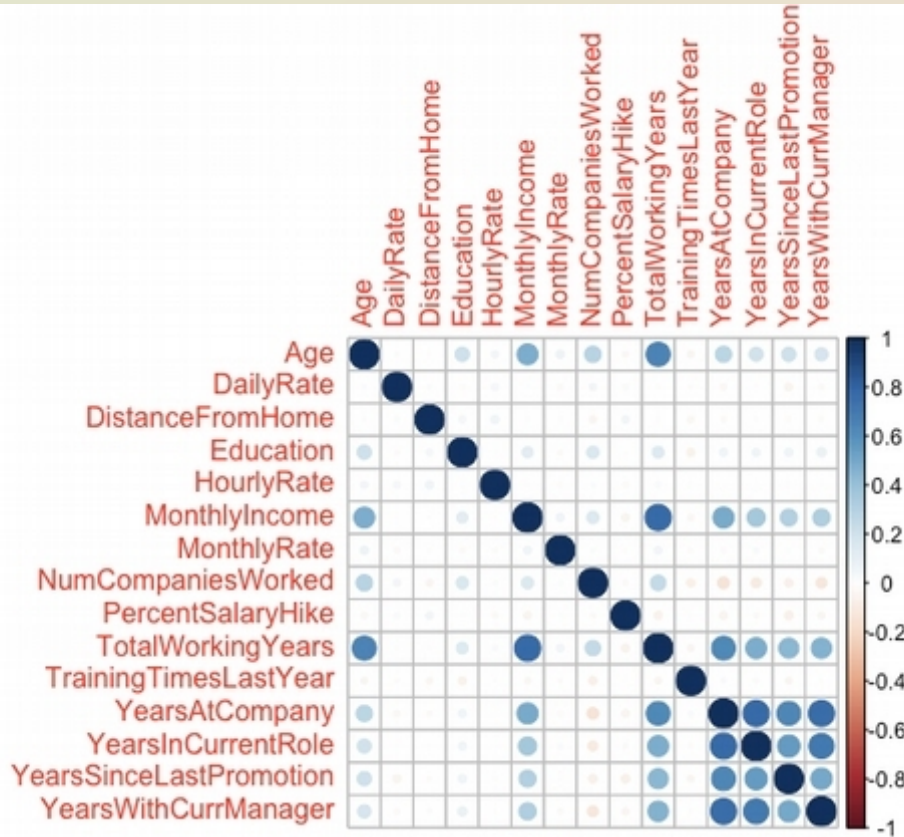
Attrition Distribution



MonthlyIncome Distribution



EDA continued – Numeric Correlations



The numeric correlations don't have any strong negative relationships.

The years derived features share relationships.

As a whole, the strong relationships are intuitive.

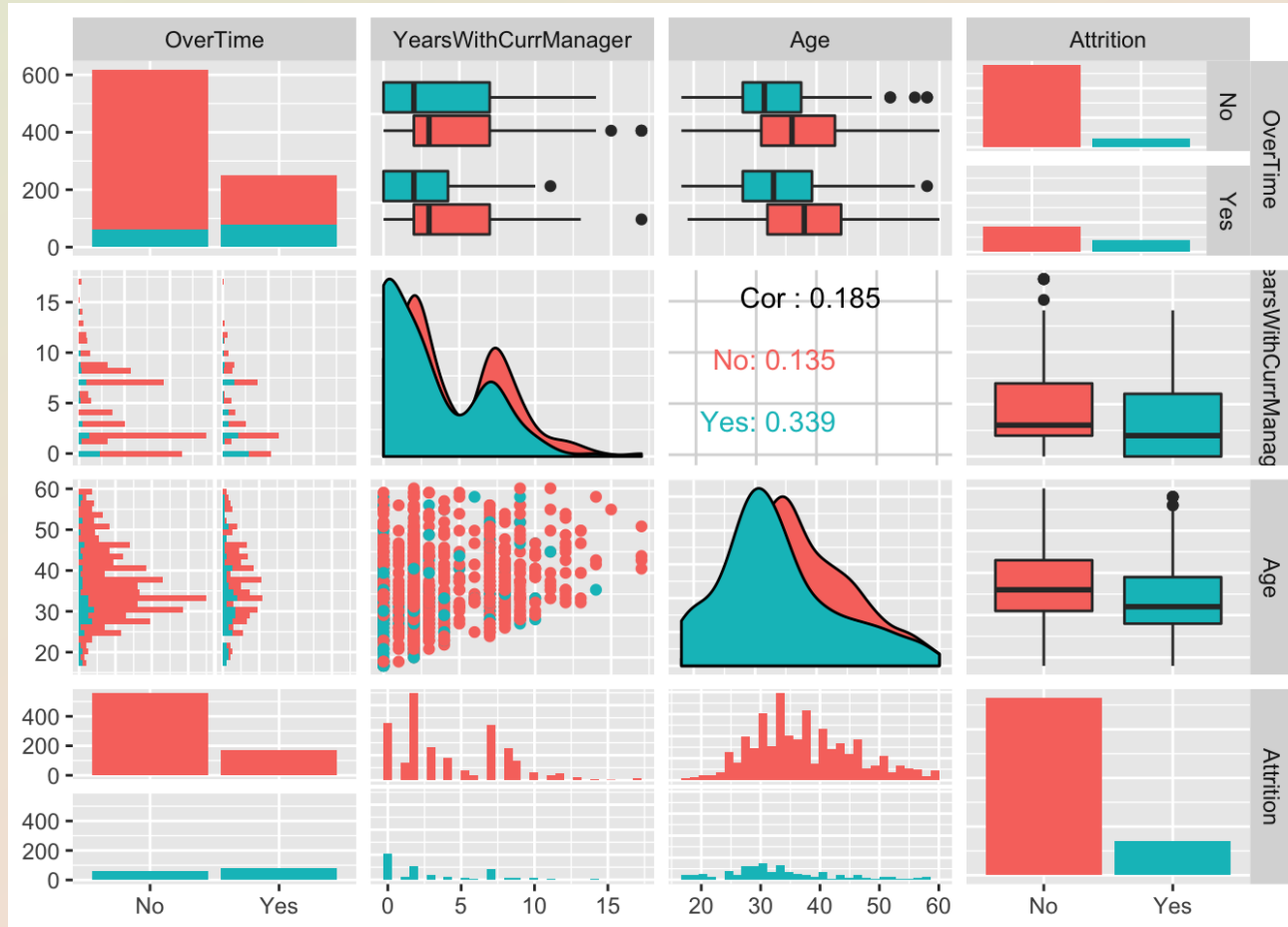
Feature Selection

- We used several algorithms for top feature selection (most important/predictive)
 - Multivariate Adaptive Regression (MARS)
 - Random Forest
 - Step-wise Regression

Feature Selection - Attrition

- Top Three Predictors for Attrition:
 - Overtime
 - Years With Current Manager
 - Age

Feature Selection – Attrition Visual

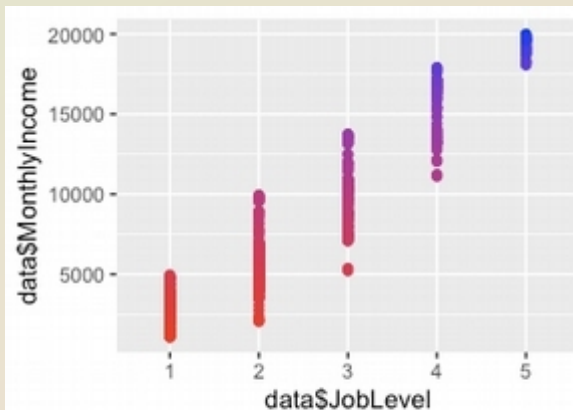


Feature Selection – MonthlyIncome (Salary)

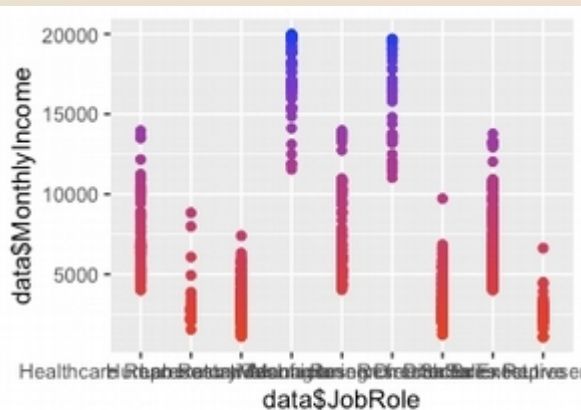
- Top Three Predictors for Monthly Income:
 - Job Level
 - Job Role
 - Total Working Years

Feature Selection – MonthlyIncome Visual

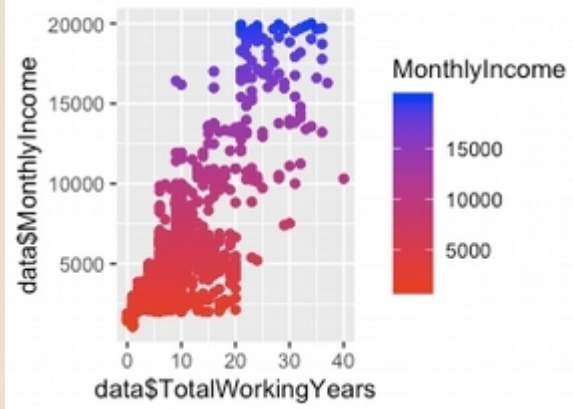
MonthlyIncome
vs JobLevel



MonthlyIncome
vs JobRole



MonthlyIncome
vs TotalYearsWorking



Modeling - Attrition

- The naive Bayes performed better than k-nearest neighbor (kNN)
- Accuracy ~84%
- Sensitivity ~90%
- Specificity > 60%

Modeling - MonthlyIncome

- The best performing model utilized multiple linear regression (MLR) on the top three predictors previously mentioned
- Root mean squared deviation (RSME) is less than \$1k
- Better than MLR with all features used as predictors

Predictions Provided

- Using our best performing models, we labeled the non-labeled datasets provided by Frito Lay.
- Labeled prediction datasets in “prediction” folder deliverable

Conclusion

- We hope Frito Lay finds our models useful in their ongoing operations for employee attrition and salary prediction efforts
- Final Models used:
 - Attrition: naive Bayes
 - Monthly Income (salary): Multiple Linear Regression