Predicting Attrition- A Driver for Creating Value, Realizing Strategy, and Refining Key HR Processes

Kevin Mendonsa, Maureen Stolberg,
and Vivek Viswanathan

SMU
DataScience@SMU

Introduction

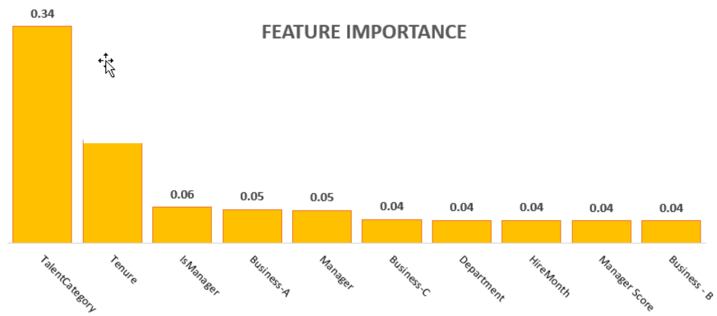
- **Strategically managed attrition is good as "Human Capital", comprises 70 -90% of an organization's cost**
- **Talent is the key competitive differentiator today.**
- **Retention and new talent is the key to growth, innovation, and success.**
- **Attrition and turnover are key performance metrics for any organization.**
- **Understanding the drivers of attrition are crucial to address underlying organizational issues and develop remediating strategies.**

Problem

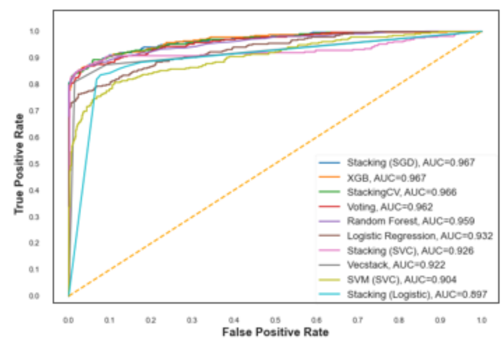
- **Predict Attrition by leveraging established machine learning techniques and relevant models that are most applicable to a specific organization**
- **Identify the major factors that contribute to attrition within a firm**
- **Provide insights into key HR processes that potentially impact employee attrition**

Feature Importance

- **Talent Assessment Rating:** The assigned category for every employee and importance to the firm
- **HireYear:** This is synonymous with tenure of an employee within Company "X" was also deemed important.
- **Allocation to a specific business or division:** The specific business that an employee worked within
- **Employee Category (Manager/Employee)**
- **Employee's Manager and department (CCID)**
- **Manager Engagement Score**



Best ROC Curves by Classifier



Classifier	Accuracy	Recall	Precision	AUC Score
Stacking (SGD)	0.929	0.868	0.913	0.967
XGB	0.929	0.868	0.915	0.967
StackingCV	0.923	0.863	0.9	0.966
Voting	0.932	0.857	0.932	0.962
Random Forest	0.935	0.83	0.969	0.959
Logistic Regression	0.866	0.798	0.798	0.932
Stacking (SVC)	0.927	0.854	0.92	0.926
Vecstack	0.898	0.874	0.826	0.922
SVM (SVC)	0.861	0.78	0.796	0.904
Stacking (Logistic)	0.894	0.836	0.842	0.897
Bernoulli Naive-Bayes	0.815	0.791	0.694	0.889
Decision Tree	0.883	0.868	0.798	0.879
Gaussian Naïve-Bayes	0.524	0.982	0.409	0.655

Model Approaches and Evaluation

- **Data pre-processing:** Data Imputation, *one hot encoding*, Scaling, Checking for correlation, up-sampling (SMOTE)
- **Feature reduction:** Feature ranking with recursive feature elimination and cross-validated selection of the best number of features
- **Model Techniques:** Logistic Regression, SVM, Naïve Bayes, KNN, Decision Tree and Random Forests, Ensemble Models (Vecstack, StackingCVClassifier, StackingClassifier, VotingClassifier)
- **Parameter Selection:** Randomized search on hyper parameters – RandomizedSearchCV.

Model Effectiveness

- **Validation –** Train-Test split, Cross-validation
- **Effectiveness –** Accuracy, Recall, Precision and ROC scores
- **Aim of analysis –** Reduce False Negatives, Improve recall score
- **Inferences –** Gaussian Naïve Bayes had the least false positives but a huge false positive classification. Ensemble models were performing better than individual models. Performance of Random Forest and XGBoost models were comparable to the Ensemble models
- **AUC Scores:** Vecstack, StackingClassifier, VotingClassifier, and RandomForest all had AUC score > 0.95

Conclusions

- Helps convert a "lagging indicator" into a leading one.
- The predictive models provide key insight into which factors are important as levers to control and manage attrition. If properly used, can be invaluable to any business looking to proactively manage attrition.
- Talent, Tenure and Employee type clearly are the most important factors driving attrition at Company "X" in addition to Manager Engagement scores and Business.
- It is recommended that Company "X" include additional factors to determine what other levers are available to address attrition proactively.