Claims Classification Project | Milestone Seven

Executive summary prepared for Salifort Motors leadership by the Salifort Motors data team

ISSUE / PROBLEM

Salifort Motors leadership and its HR department asked to develop a model that will help them better understand their employees and improve employee retention. In this part of the project, the data team needs to create a machine learning model that predicts whether or not an employee will leave the company.

RESPONSE

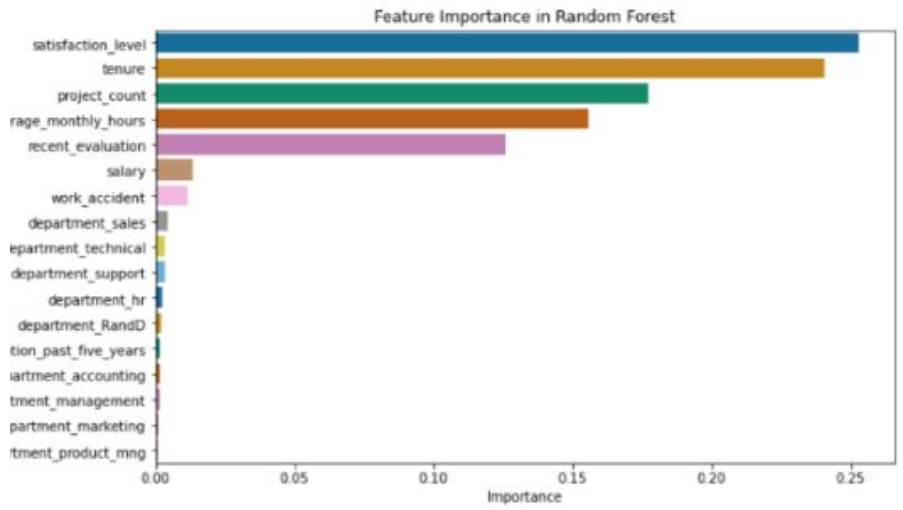
The data team chose to develop two different machine learning models (i.e., random forest and XGBoost) to cross-compare results and obtain the model with the highest predictive power (i.e., random forest).

Seeing the confusion matrix, the model correctly predicts true negatives and true positives; yielding almost no false negatives or false positives.

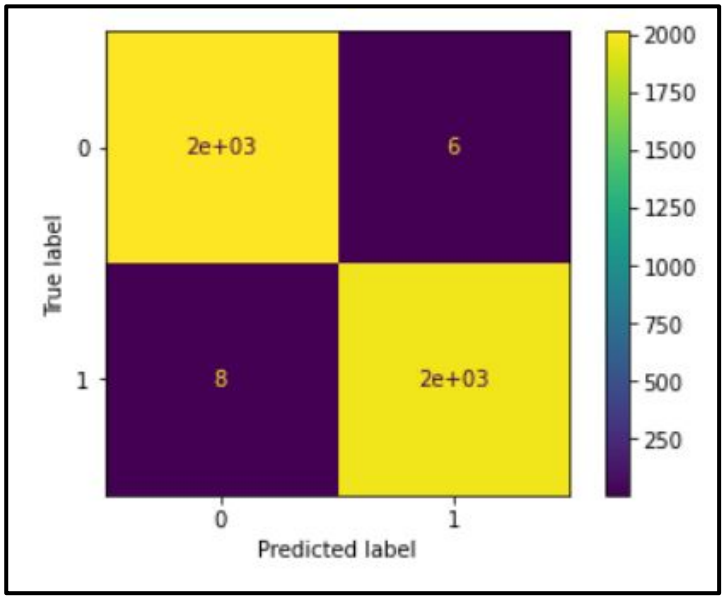
IMPACT

Developing tree-based models since tree-based models are robust to outliers or values that deviate significantly from the norm.

The machine learning model provides an exceptional framework for predicting binary outcomes indicating whether or not an employee will leave the company.



The bar plot above shows the relative feature importance of the predictor variables in our machine learning model.



The upper-left quadrant displays the number of employees that the model accurately classified as staying. The lower-right quadrant displays the number of employees that the model accurately classified as leaving. The upper-right quadrant displays the number of employees that the model misclassified as leaving. The lower-left quadrant displays the number of employees that the model misclassified as staying.

KEY INSIGHTS

- **The bar plot of the relative importance of the features in the model indicates the most predictive features were related to satisfaction levels and the workload of employees**. The top five important features in the machine learning model are “satisfaction_level”, “project_count”, “tenure”, “average_monthly_hours”, and “recent_evaluation.”
- **After conducting feature engineering and cross validation,** we see the Tuned Random Forest model performed exceptionally well. The model achieved a precision of 99.3%, a recall of 99.1%, an accuracy of 99.3%, an f1 score of 99.3%, and an AUC of 99.8%.
- **The data team recommends building a K-means model on this data and analyzing the clusters**. Plus, we may optimize our models using PyTorch.
- **The model and the features importance extracted from the models confirm that employees at the company are overworked**. This seems consistent with the scatter plot visualizations in the jupyter notebook showing people who left noticeably increased when "average_monthly_hours" surpassed approximately 240 hours per month.