**Dataya Ait Bulgular**

- Öncelikle datamızı model için hazırlarken hiçbir rowda NaN değer olmadığını görüyoruz.
- Bazı featureların(örn. Over18) tek bir değere sahip olduğunu ve bunları modele hiçbir etki yaratmayacağı ve
- sadece uzayı büyüteceği için kaldırıyoruz.
- Bazı featureların(örn. OverTime) sahip oldukları değerlerin one hot encoding yapmadan önce uygun sayısal
- değerlere map ettik. Örneğin OverTime yes:1, no:0 şeklinde. Bunları one hot encoding ile de yapabilirdik
- ancak mapping yaparak uzayın boyutunun daha da büyümesini engellemiş olduk.
- EmployeeNumber sadece employeelerin sahip olduğu unique numberları içerdiği için ve feature olarak bize hiçbir şey katmadığı için datamızdan onu da kaldırdık.
- Joblevel ve MonthlyIncome columnları arasında korelasyon değeri 0.95'ten büyük olduğu için ve Income bize
- daha geniş bir aralık sunduğu için JobLevel'ı feature olarak kaldırdık.
- Bunlara ek olarak dublicated row var mı diye, datamızı aynı değerlerle iki kez eğitmemek için, kontrol ediyoruz ve olmadığını görüyoruz.

**Modellere Ait Bulgular**

- En uzun sürede train edilen model Logistic Regression, en hızlısı ise Random Forest.
- En uzun sürede optimize edilen model Random Forest, en hızlısı Logistic Regression.
- Datamızda 0-1'lerin dağılımı dengesiz olduğu için başta eğittiğimiz modeller daha fazla olan 0'ları doğru tahmin etmeye daha da meyilliydi. Accuracy değerleri yüksek olsa bile recall değerleri oldukça düşüktü. Ancak hyperparameter optimizasyonu ve Logistic Regression ile Random Forest üzerinde threshold optimizasyonu yaptıktan sonra modellerimizin Recall ve bununla birlikte Weighted F1 skorlarının oldukça arttığını görüyoruz.
- En iyi model olarak Optimized Logistic Regression seçilmiştir. Sebebi ise Recall değerinin en yüksek olmasıdır.
- En iyi beş feature olarak ['Age', 'MonthlyIncome', 'OverTime', 'TotalWorkingYears', 'YearsInCurrentRole'] bulunmuştur.
- Optimized Logistic Regression modelimizin bu beş feature ile fit edilip X_validation üzerinde yaptığımız testlerde ise
- Recall değerinin arttığını ancak precision ve accuracy değerleriyle birlikte weighted F1 skorlarının azaldığını gözlemliyoruz.
- Bunun sebebi de Logistic Regression'ı optimize ederken tüm 47 feature'ı kullandık ve ona göre optimize ettik ve sadece beş feature ile fit edince modelimiz eskisi kadar iyi çalışmadı.
- Optimized logistic regression modelimiz pickle olarak kaydedilirken, threshold değerinin de kaydedilmesi için [logistic_regression_model, optimal_threshold] şeklinde model.pickle kaydedilmiştir.