**Antonio Serrano Redondo**

**Churn Projekt**

**LESSONS LEARNT**

**Modell Inhalt**

Was ich am besten gelernt in diesem Projekt habe sind die verschiedene Schritte zu nehmen um ein Final Score zu haben. Der GridSearch war mir nicht ganz klar. Ich dachte es war ein Art von Column Importance aber habe ich jetzt viel besser verstanden, dass die letzte Schritt für einen optimierten Score ist.

Bei Modell, habe ich auch gemerkt, dass Columns Importance kann eigentlich zweideutig sein.

Zum Beispiel hatte ich vor der Code genau schreiben ein bisschen mit der Data mitgespielt. Ich hatte bei verschiede Spalte gefiltered und die mean von Y Spalte kalkuliert. Leider hatte ich keine große Unterschied zwischen die Mean() von Event = 1 und Event = 0. Allerdings bei Profession Commercial gefiltered, hatten in Event 1 die mean() = 7 und bei Event = 0 mean() = 5. Das fand ich eine gute Anmerkung. Ich kann mir vorstellen, dass die Leute die in Commercial arbeiten , haben Sales Ziel jeder Monat und das Gehalt hängt davon ab, deswegen die Leute die mehr Anxiety haben, können damit nicht umgehen und am Ende kündigen.

Deswegen dachte ich, dass es für den Baum Erstellung Profession wichtig sein könnte, allerdings war die Spalte Profession rausgenommen. Das war nur ein kleines Beispiel aber sicher dass es viel mehr gibt, die ich nicht gesehen habe, die raus von Modell wären. Daher habe ich gelernt, dass Decission Tree besser mit spezifische Information arbeiten und nicht mit zu viele Spalten, weil der Baum zu groß wird. Ich frage mich, ob es gibt die Gelegenheit die Importance Column nutzen anders: Nicht um Columns rausfiltern sonst für Spalte-n wie Wurzeln benutzen, um der Baum große zu reduzieren.

Ich habe auch gelernt, wie wichtig ist eine gute Kategorie Verteilung erstellen in der Daten Vorbereitung. Mann muss nicht nur die Strings in Zahlen umsetzen. Die Zählen Gruppierung sollten eine Verbindung mit der Realität haben. Ein Beispiel in unsere Gruppe war die Gruppierung von Spalte “Way”. Wir haben By Foot, Car, Bus. Die Gruppe hat es interpretiert wie der Distanz von der Arbeit und dachten, dass eine angemessene Gruppierung wäre By Foot = near, Bus = middle, Car = far. Ich dachte dass es besser wäre, nur 2 Gruppe machen: By Foot = Without vehicle und Bus - Car = With vehicle (Meine Grund war, dass vielleicht die Jungen unabhängig von der Distanz den Bus einfach nehmen, weil sie kein Geld haben, um ein Auto zu kaufen). Beide Interpretationen der Daten sind gut aber jeder Option hat ein Einfluss in der Finale Baum. Der Lesson ist: Die Interpretation der Features in der Realität, dass wir analysieren möchten, hat eine sehr große Einfluss in der Ergebnisse. Deswegen lohnt es sich, genug Zeit nehmen zu überlegen in der Daten Vorbereitung.

Was ich leicht fand ist ein Score zu bekommen. Es ist relativ einfach mit dem Library einen Score zu rechnen. Allerdings ist es sehr schwierig, der Score zu optimieren. Ich habe auch gemerkt, dass es sehr hilfreich wäre, besser die Parameter Inhalt zu lernen. Daher hätte man mehr Gelegenheiten, es zu optimieren.

Es war auch sehr schwierig der Projekt ausführen in 3 Tage, vielleicht nicht genug Zeit dafür. Insbesondere für Leute ohne Erfahrung in Machine Learning.