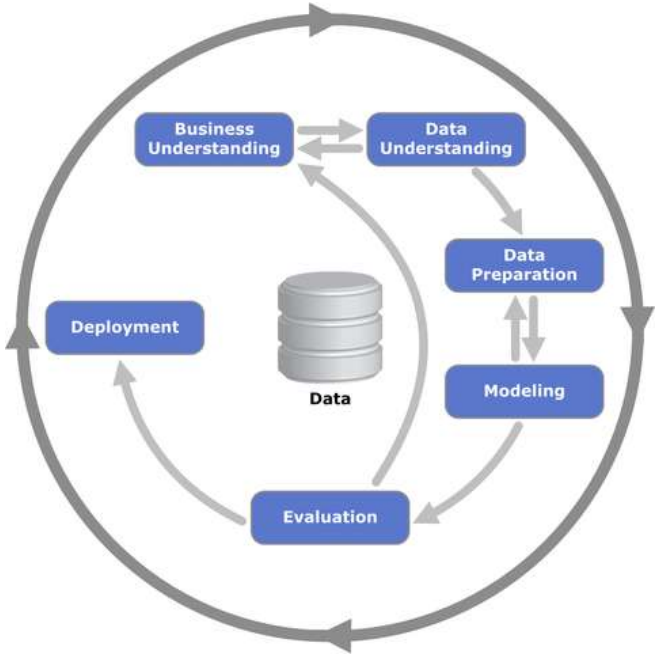
Syriatel Customer Churn

Crisp DM Checkpoint #2

Vivian Dang and Joe Buzzelli

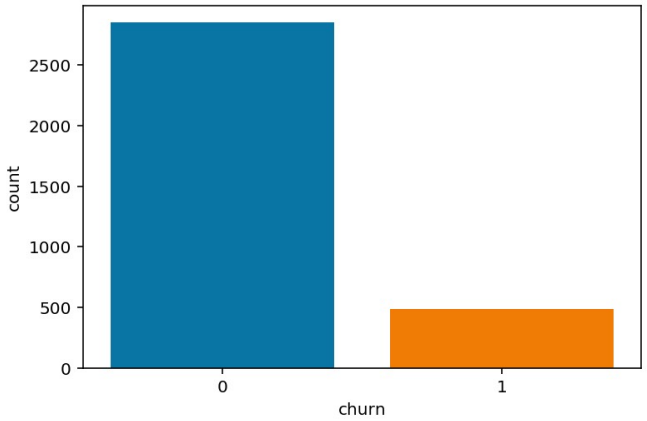


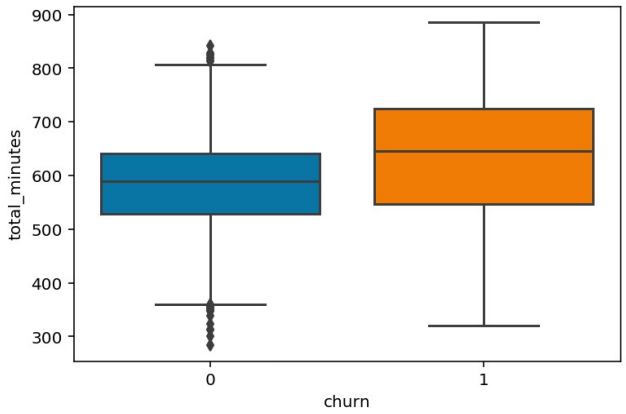
April 14, 2020

EDA/Visualization

As of 04/13, created EDA visualizations exploring the baseline data pertaining to customer churn

- Baseline is roughly 14% churn
- Total minutes most significant delta identified between classes





We are pursuing a classification modeling strategy

Our goal is to predict customer churn by creating a model that maximizes **precision** with a target greater than or equal to **0.95**

Models used include:

- Logistic regression
- Decision tree
- Random Forest
- Nearest Neighbor
- Boosting

In order to mitigate overfitting, we will conduct **cross validation** and **hyperparameter tuning** on our two best performing models

- Additionally, we are exploring feature selection by leveraging the best feature calculations available in decision tree and random forest models

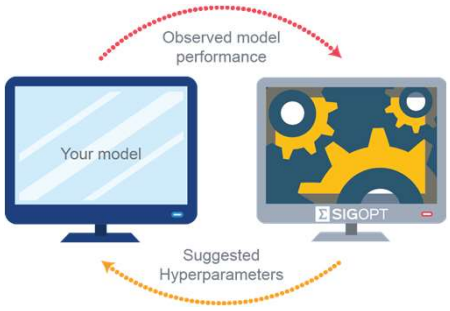
Initial Model Results (perfect scores equal -1.000)

Model	Precision	Accuracy	Recall	F1
Decision Tree (Tuned, max_depth = 5)	**0.979**	0.858	0.997	0.92
KNN (k=4)	**0.978**	0.904	**0.343**	**0.51**
Logistic Regression (default w/ SMOTE)	**0.952**	0.925	0.896	0.92
KNN (default)	0.942	0.913	0.429	0.59
Random Forest w/ SMOTE (max_depth=4, n_est=5)	0.9273	0.928	0.879	0.92
KNN w/ SMOTE (default)	0.728	0.669	1	0.84
Logistic Regression (default)	0.704	0.883	0.331	0.45
Decision Tree (max_depth = 5) w/ SMOTE	-1	0.916	0.832	0.91
Random Forest (default)	-1	-1	-1	-1
Random Forest (max_depth=4, n_est=5)	-1	0.888	0.228	0.37
Decision Tree (default)	-1	-1	-1	-1

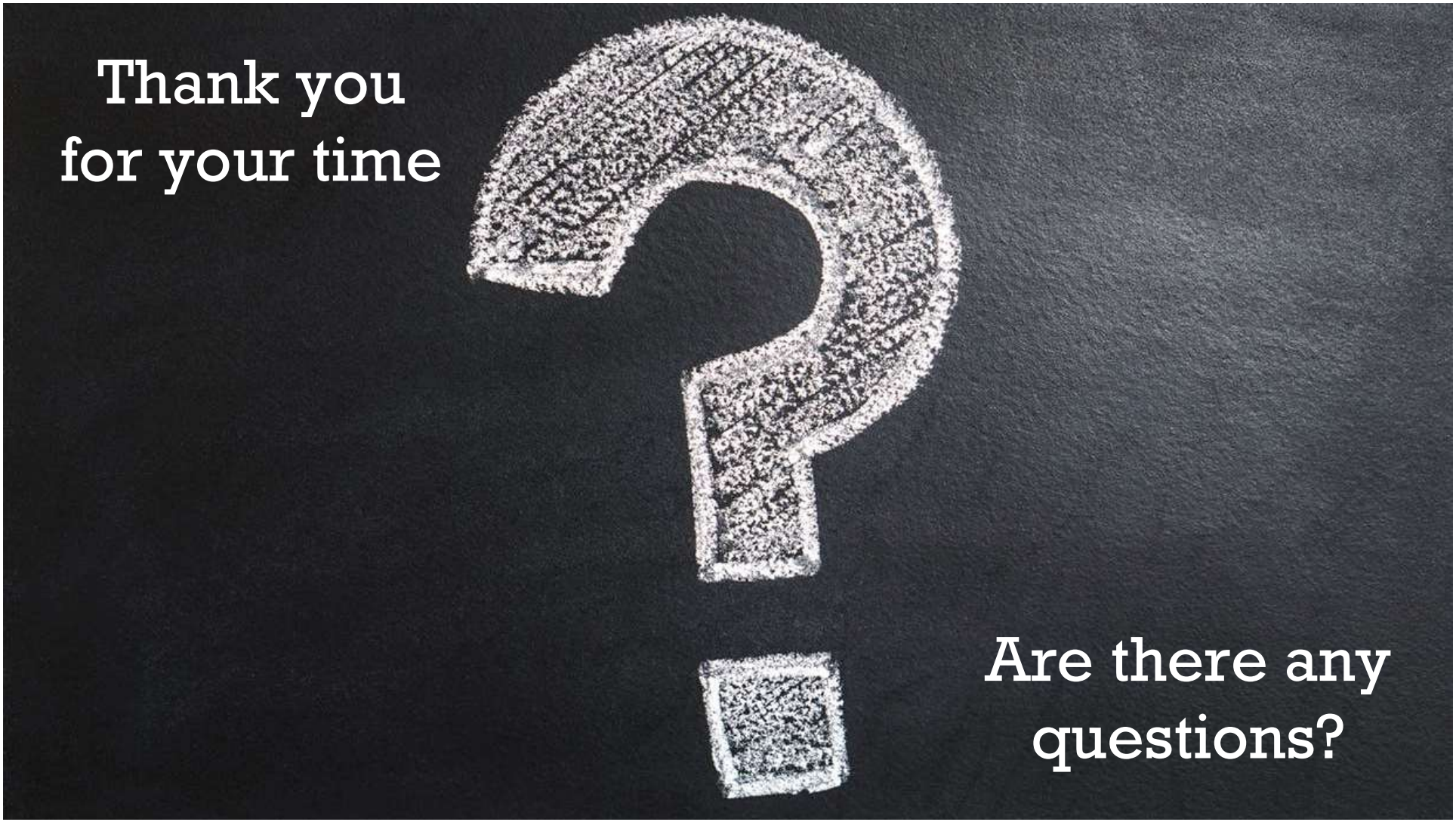
We are pursuing a classification modeling strategy

In order to mitigate overfitting, we will conduct **cross validation** and **hyperparameter tuning** on our two best performing models

Additionally, we are exploring **feature selection** by leveraging the **best feature** calculations available in decision tree and random forest models



Thank you
for your time



Are there any
questions?