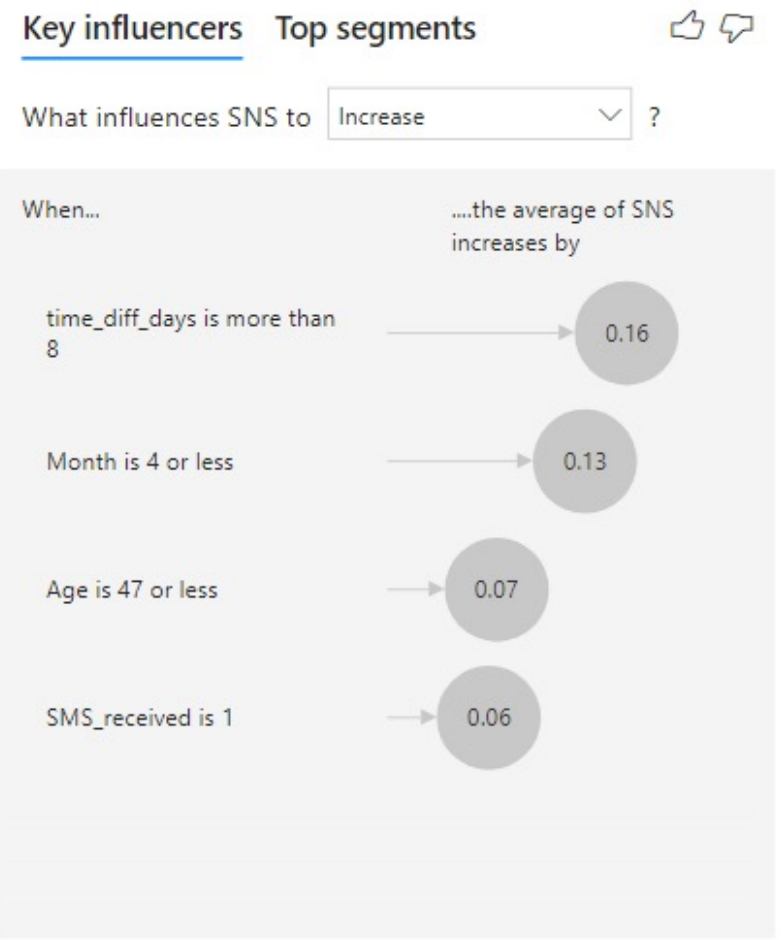
Hospital Appointment Show-No Show Prediction

09/03/2021

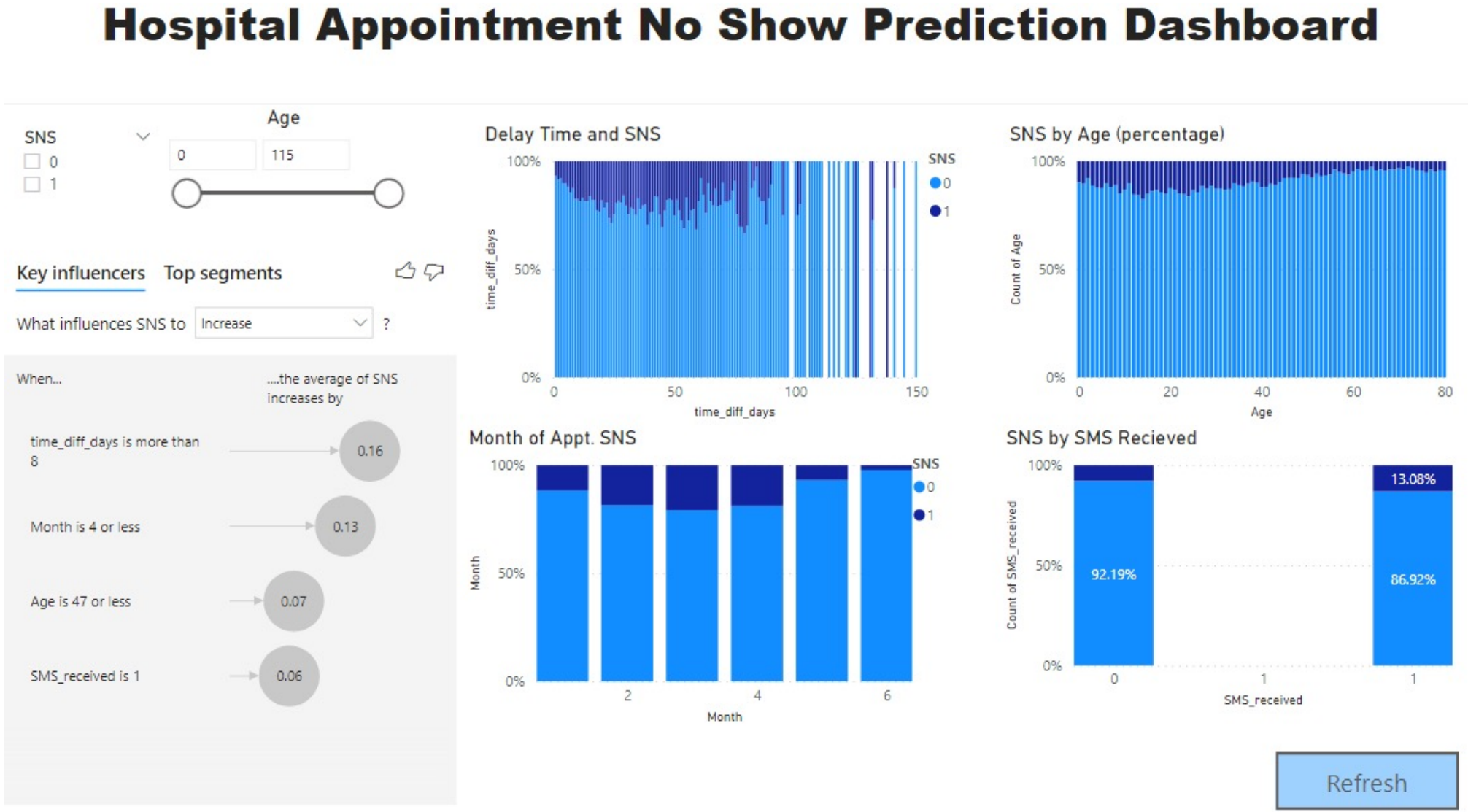
Hong Tang

Summary and Recommendation

- **Why:** No show causes profit loss and resources wasted
- **Findings:** Following features are good indicator of Hospital Show/Noshow (SNS), Right figure
- Randomforest Classifier outperforms other classification methods; final selected model has robust AUC 0.72 for both training and testing set
- **Recommendation** to improve show prediction
  - Focus on efforts on patients with longer delay time
  - Survey on patients with no show
- **Plan Forward:**
- Improve feature engineering; Investigate interaction of features; Model tuning



Preliminary Dashboard/Mobile App



ML Process

Cycle 1

Feature EDA

Existing numerical features

Model Selection

Simple Logistic regression

Cycle 2

Feature EDA

Time Series features

Numerical features
Categorical features

Feature importance:

Three ranking methods

Model Selection

4 Classifier benchmark

Cycle 3

Further feature engineering

Finetune RFC

Model Deployment

Model Deployment

Dashboard and web App (ongoing)

Key Features

Feature Explored

- Weekday for schedule
- Weekday for appointment
- **Time difference between schedule and appointment day (or “Delay time”)**
- PatientID
- **AppointmentID**
- **Month** of schedule day

- Note: high collinearity between appointmentID and Delay Time, appointmentID is removed from model building, which improve model prediction accuracy

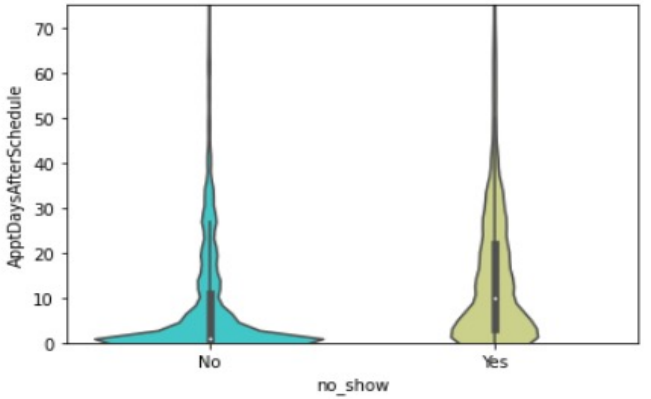
No Show indicators

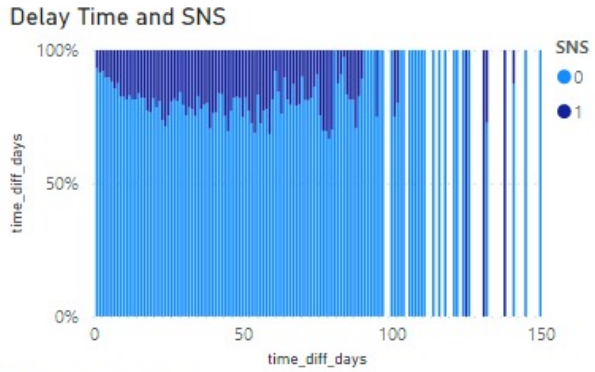
Age 20-60 “busy working” group

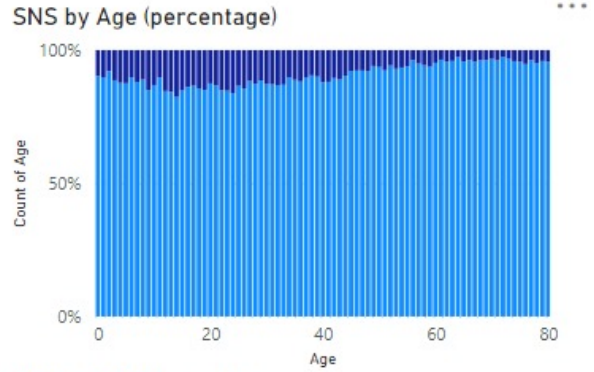
Longer delay time

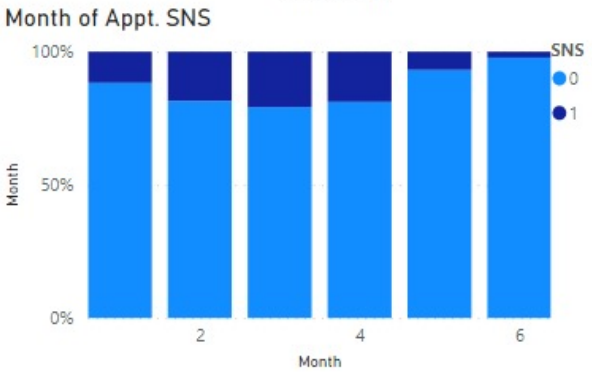
January-April and

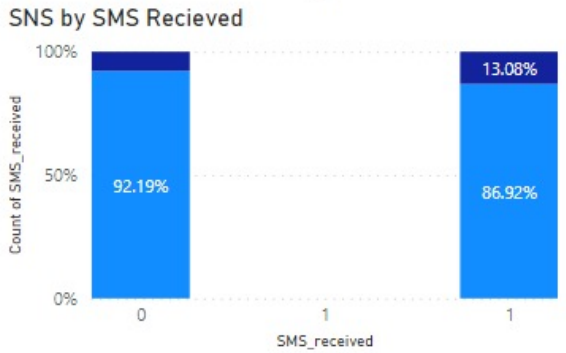
Received message





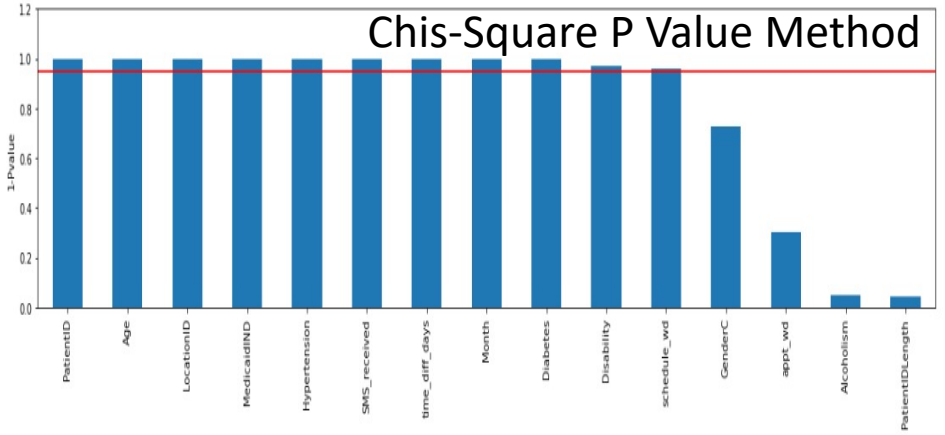


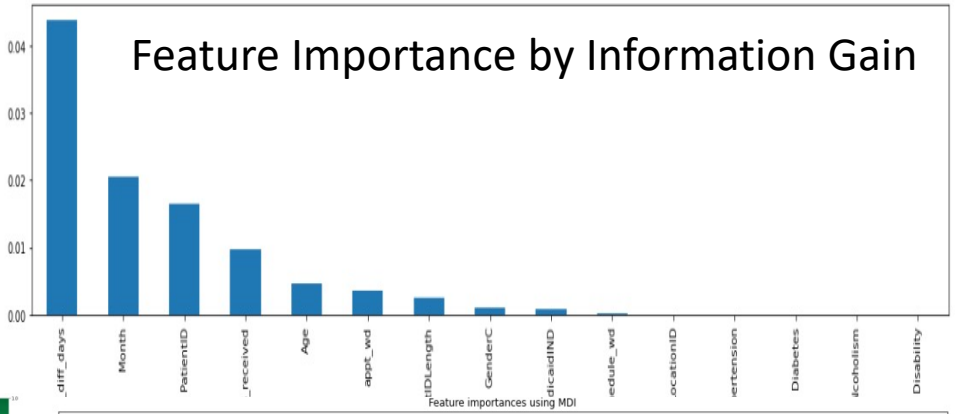


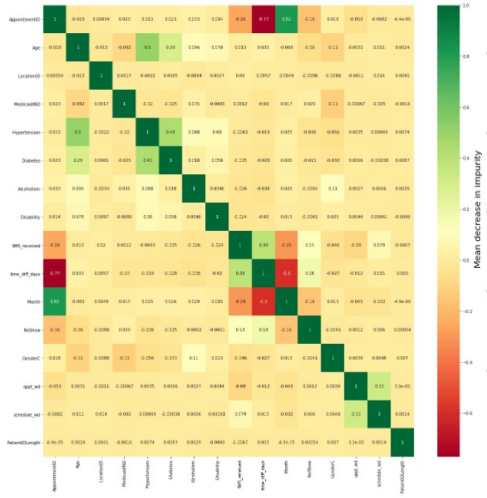


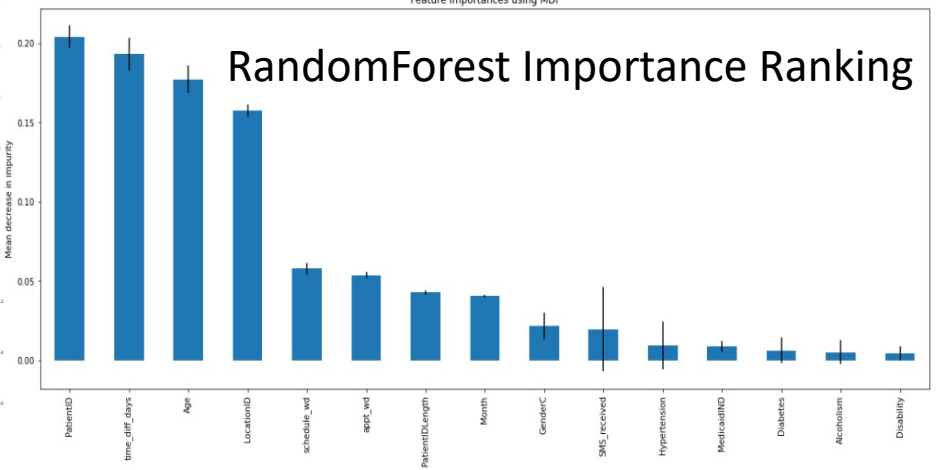
Feature Selection

- Three methods are used to select features along with correlation matrix
- The time difference between schedule and appointment is the most important feature for SNS
- Location ID and Age are also important for SNS prediction









Model Selection and Hyper Parameter Tuning

Four Classification Methods are used for model selection

- **Random Forest**
- Logistic Regression
- Naïve Bayes Gaussian
- Decision Tree

RandomForest is selected for prediction

	model	best_score	best_params
0	random_forest	0.799116	{'max_features': 'auto', 'min_samples_leaf': 2...
1	logistic_regression	0.796789	{'C': 1}
2	naive_bayes_gaussian	0.796789	{}
3	decision_tree	0.720595	{'criterion': 'entropy'}

Grid based search on three most important factors for RFC

- N_estimators
- max_features
- min_sample leaf