# Executive Summary

## Business Understanding

With SAS’s environmentally responsible mindset the company plans to be sustained completely by clean energy. They have managed to cut global emissions across all scopes by decreasing them by 37.7% in 2020. They have also generated 3.8 million kWh of clean, renewably sourced energy from rooftop and ground-mounted solar systems, and are committed to 2050 net-zero carbon emissions. Their headquartered out of Cary, NC where these positives are seen implemented on its campus.

SAS has reached out to us to solve a unique problem at their headquarters branch. As they aim to rely fully on solar energy, the company must find ways to evaluate energy demand and consumption powering Building A on the site. In previous insight we have determined that it is more economically efficient for SAS to focus efforts on adding more panels to the farm opposed to implementing a battery system outlined in [phase one](file:///S:\zpdenton\MABA\Practicum\Technical%20Report.docx) of the project. I have constructed several models to predict their solar farm output. This model can be used to forecast costs and quantity of additional solar infrastructure. Additionally, create an interactive dashboard that will allow them to view what the solar output will be for any given day of the year based on historical data.

## Interactive View

This interactive dashboard allows the user to view significant metrics related to solar farm output. The user can set the dashboard to any month of the year to view the solar output along with the measure of values for the respected metrics for that month (averaged over the years).

## Analysis

After the data was properly prepared a final modeling pipeline was created. This pipeline included 11 models. The ensembled model produced the lowest error of 1041.393. This means our solar prediction can be wrong by up to 1041.393(kWh) from the actual solar output for a given day. When comparing this to the average solar output of 4001.57(kWh), the model can mis-predict solar output kWh up to 26% of the time. While this is a moderate amount of error, SAS can predict 74% their solar farm output. I believe this still gives SAS viable benefit in predicting. I recommend SAS should consider exploring ways to reduce this error by collecting additional data on building A specifications, and historical UV/thermal data over Cary, NC.

## Deployment

Being able to predict solar production for the next day will allow SAS to determine how much energy is being pulled from the grid. SAS will then be able to derive future costs of additional energy. Furthermore, they will be able to know how allocate additional panels to cover grid expenses. The model can also be implemented in other SAS locations with solar farms for the same purposes. With this analysis, SAS can efficiently balance energy demand and consumption at their headquarters.

# SAS Solar Farm Predictive Analysis

## Vision Statement

**My vision is to develop a predictive model for SAS that will enable the company to predict the energy consumption levels for their HQ building “A”, accounting for a variety of weather conditions**. **In addition, create an interactive dashboard for management to view the buildings consumption in these different conditions.**

## Project Scope

The final product will be several predictive models that uses weather, consumption, and feature engineered data that will predict building A’s solar farm output on any given day. This will in turn identify how to effectively power the building given specific conditions and establish excess/deficit solar production.

## Data Management

Datasets, visualizations, and models are stored in a dedicated directory for the SAS solar project. Given that the data provided is static, and the scope of the project new data will need to be updated manually. The appendix contains views of the [data dictionary](#dataDictionary).

## Data Preparation

Data cleaning and transformation steps are as follows:

* Weather data: Excel
  + Split date/time columns to have date column.
  + Split hourly sky conditions.
    - Do this for low, med, and high altitudes.
    - Remove split columns without this data.
  + Created dummy variable for hourly sky conditions.
    - Splits should be made to remove all text to display only the three-character nominal sky condition (ex: BKN).
  + Created a week/weekend nominal variable.
  + Created daily precipitation nominal variable (no, yes) to indicated whether there was precipitation or not for that day.
  + Created column to calculate daily sunlight time in hours.
    - Subtracted sunset time from sunrise to get daily sun light.
    - Converted sun light time from military time to hours (100ths).
  + Transformed all columns to daily totals.
    - Create daily totals or averages for records.
  + Dropped unused columns that are not in data dictionary.
* SAS data: Excel
  + Transformed SF1 (solar farm), bldg. A demand to daily totals
  + Combined and matched corresponding dates into one data set.
* Combined data: SAS Miner
  + Variable transformation node:
    - Standardize interval inputs
    - Dummy indicator for class inputs
* Variables that are highly skewed were transformed into bins or nominal data.

## Data Exploration

Here are some insights we found from transformed data:

* We can see that solar output has a pattern across years where maximum output occurs in the late spring/early summer while minimum output occurs in late fall and winter months ([Figure](#Fig1) 1).
* Over the time span of the data, yearly maximum outputs have decreased over time ([Figure 1](#Fig1)).
* There is a correlation with average hours of daily sunlight and solar output. Months with higher output have more average hours of sunlight and months with less output have less hours of daily sunlight ([Figure 1](#Fig1)).
* As daily sky conditions move from high to low altitude, sky conditions show more cloud cover ([Figure 2](#Fig2)).
* All month’s average about few clouds (i.e. 3) for sky condition ([Figure](#Fig2) 2).
* There was a significant effect of if there was precipitation on solar farm output (kWh). Days that had precipitation averaged less solar output than days that did not have precipitation ([Figure 3](#Fig3)).
* There was no significant effect of whether it was a weekend or weekday on building a daily demand. Meaning there is not a significant difference in power demanded weather it is a weekend or weekday ([Figure 4](#Fig4)).
* There was a significant effect of if there was snowfall on solar farm output (kWh). Days that had snowfall averaged less solar output than days that did not have snowfall ([Figure 5](#Fig5)).
* January, August, and February have the highest count of daily precipitation across data time span ([Figure 6](#Fig6)).

## Modeling

The following models have been conducted and ensembled for predicting solar farm output. ([Figure 7](#ModelFlow)):

Ensemble:

* Default Linear Regression
* Backwards Linear Regression
* Forward Linear Regression
* Stepwise Linear Regression
* Decision Tree (Branches: 2, Depth: 6, ProbF)
* Decision Tree (Branches: 2, Depth: 6, Variance)
* Decision Tree (Branches: 2, Depth: 3, ProbF)
* Decision Tree (Branches: 2, Depth: 3, Variance)
* Neural Network
* Auto Neural Network
* PCA HP Neural Network
* HP Forest
* PCA Gradient Boost Model
* Decision tree + regression + HP Neural Network

From the final modeling phase the ensembled model produced the best results with the lowest root average squared error (RASE) of 1041.394 out of the models ([Figure 8](#Ensemble)). The ensembled model includes 4 linear regression models, 4 decision trees, random forest, boost and 3 neural networks. Significant variables:

* DailyDepartureFromNormalAverageTemp
* DailySunlightHours
* DailyVisibility
* DailyAverageRelativeHumidity
* DailyPresipitation\_yes
* DailySkyconditions\_highalt (1, 2, 3, 4)
* DailySkyconditions\_medalt (1, 2, 3)
* DailySkyconditions\_lowalt (1, 2, 3, 4, 5).

## Interactive Dashboard

This interactive dashboard allows the user to view various metrics related to solar farm output ([Figure 9](#Dashboard)). The user can set the dashboard to any month of the year to view the solar output along with the measure of values for the respected metrics for that month. The months are a cumulation of months together across the multiple years of the data. The red line on the graphs shows the monthly average and the green line indicated the overall average for that metric across all time. This gives the user an idea of how far off a month is from the overall average.

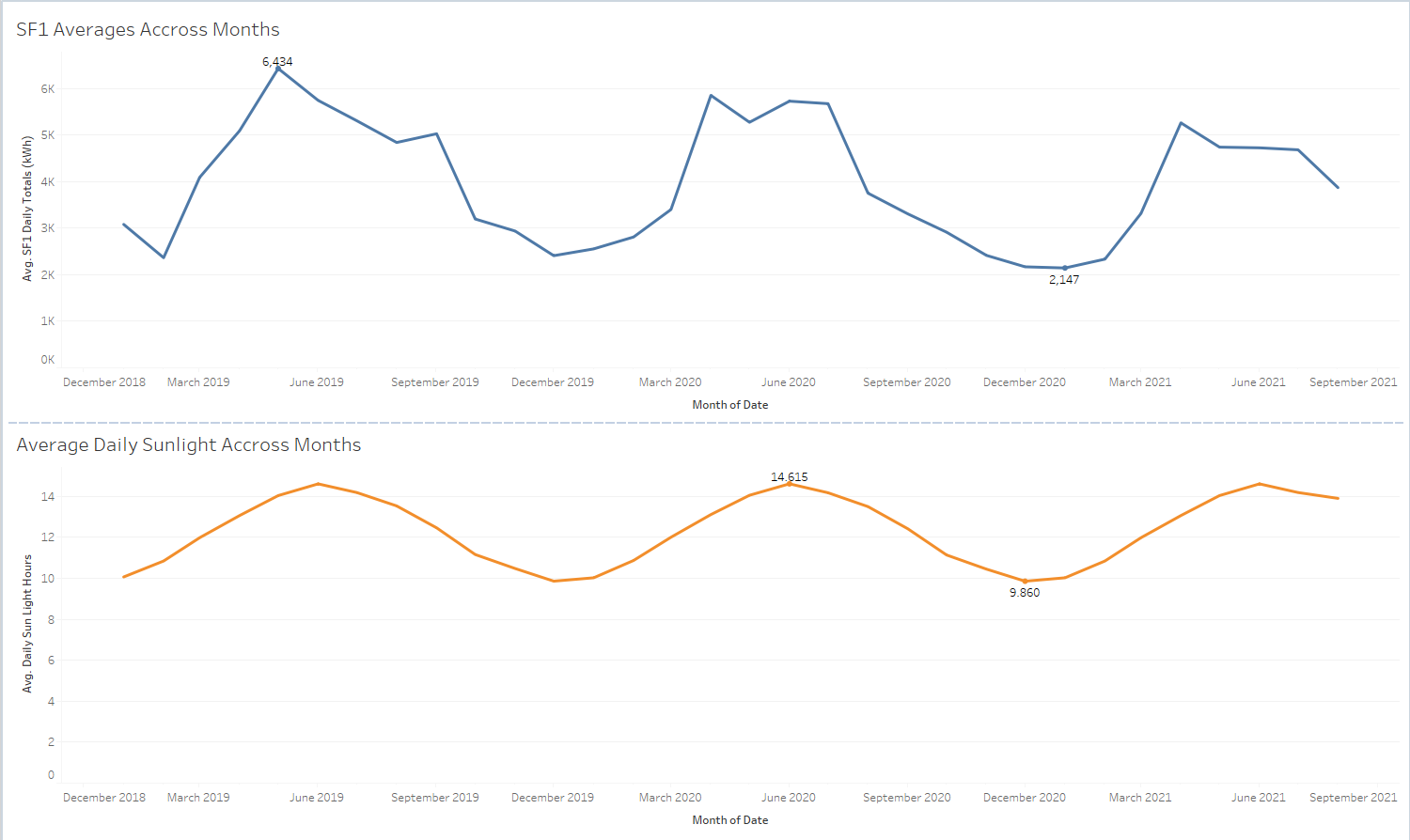
## Deployment

With this project phase the company with be able to predict solar production for the next day and allow SAS to determine how much energy is being pulled from the grid. SAS will then be able to derive costs of additional third-party grid energy. Furthermore, they will can decide how allocate financial resources to additional panels to the farm and cover grid expenses. The model can also be scaled to other SAS locations with solar farms for the same purposes. With this analysis, SAS can efficiently balance energy demand and consumption for their solar farm at their headquarters building A.

# Appendix

## Data Dictionary

## Figure 1:



## Figure 2:

## Figure 3:

## Figure 4:

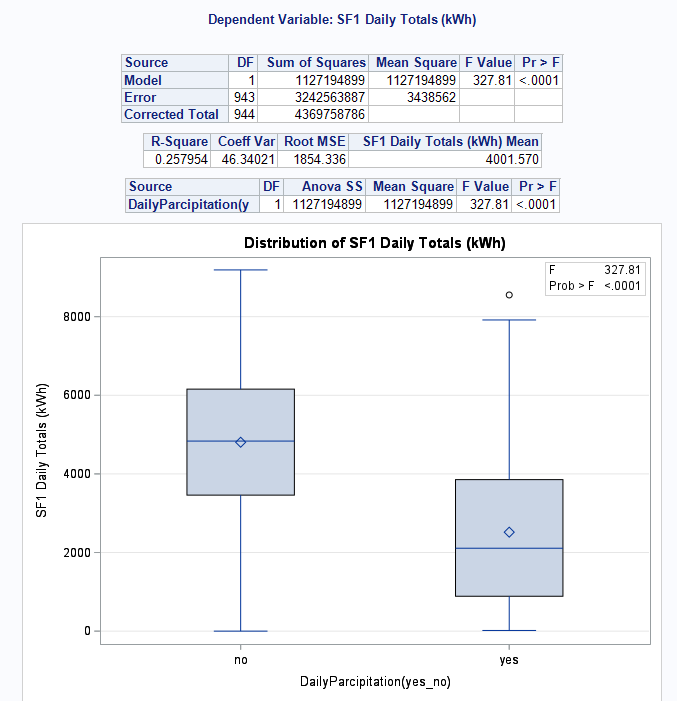
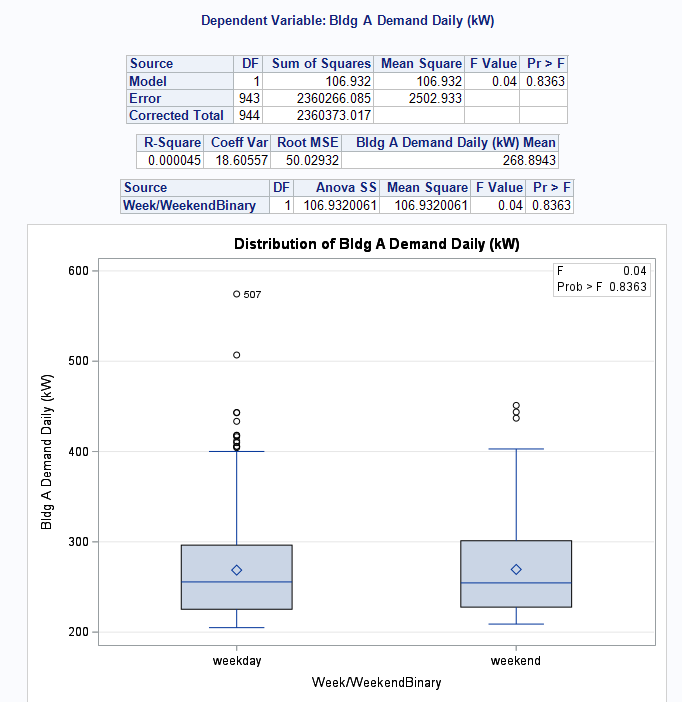
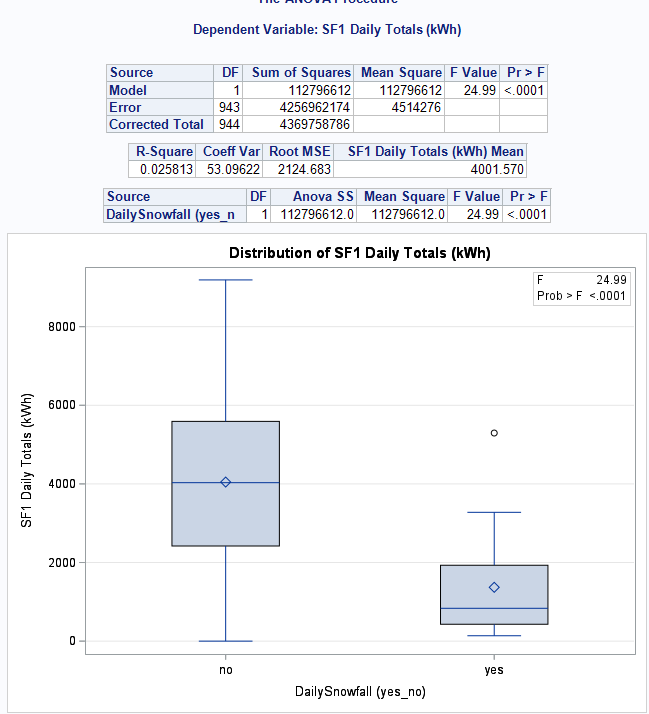


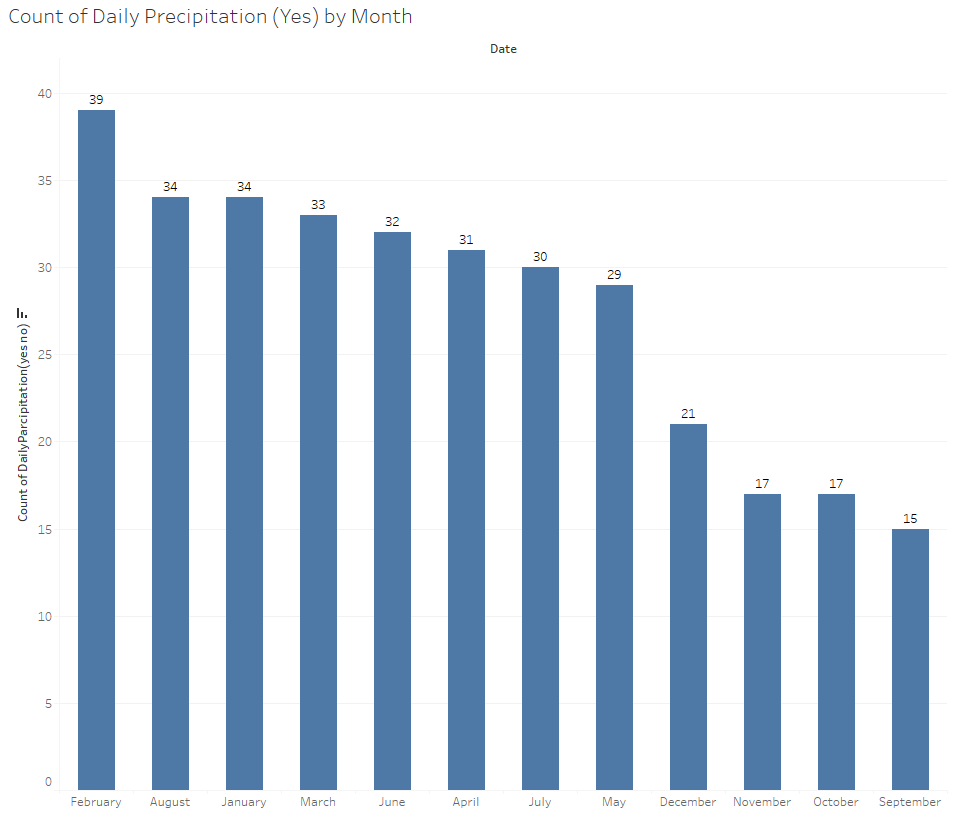
Figure 4

Figure 3

## Figure 5

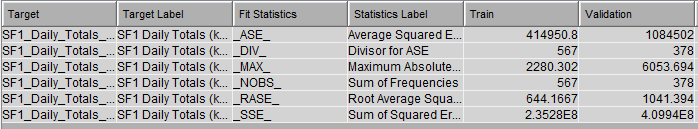


## Figure 6



## Figure 7: Modeling Pipeline

## Figure 8: Ensemble Model



## Figure 9: Interactive Dashboard