Miles Craig
February 13, 2019

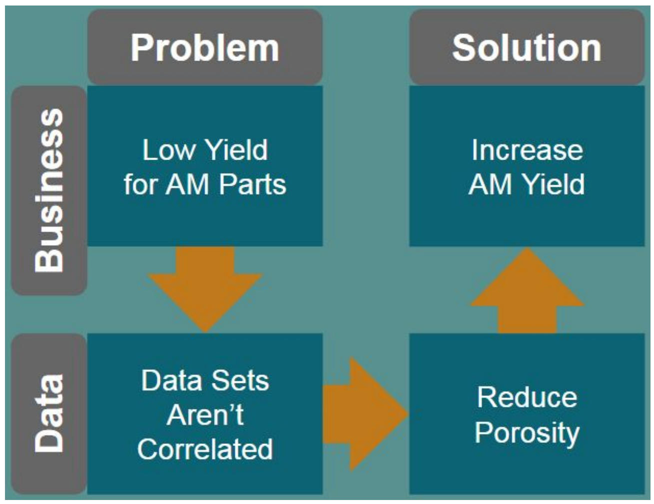
Increase AM Yield: Final Report

GitHub Project Link

Table of Contents

Problem Statement

Project Flow



1. Start with the Business Problem
    a. Low Yield of Additive Manufactured parts
    b. More parts need to be printed than order
    c. Increases the cost of time, energy, man power, delayed delivery, etc.
2. Convert the Business Problem into a Data Problem
    a. Low Yield is due to higher than desired Porosity levels
    b. Varying Porosity levels is due to the unknown correlation between Print Parameters and Melt Pool Metrics with Porosity
3. Solve the Data Problem
    a. Correlate the datasets so that Porosity can be decreased
4. Convert the Data Solution into a Business Solution
    a. Decrease Porosity on a consistent basis to increase Yield
    b. Increase Yield to reduce costs and increase profits and customer satisfaction

Business Understanding

The primary business objective is to increase the yield of metal additive manufactured parts. Producing metal 3D printed parts with Additive Manufacturing (AM) technology can be a very lucrative business. There are huge potential benefits to the manufacturing industry, however, the immaturity of the technology currently produces a part yield of approximately 30%, in other words 3 out of 10 parts are approved for quality. Therefore, a business is required to produce roughly 3 times the amount of parts to fulfill an order, increasing the operational costs and significantly decreasing the profit margin.

Data Understanding

The primary data objective is to determine which print parameters or metrics correlate best with porosity. The current state of Additive Manufacturing (AM) technology is that there is no digital thread throughout the entire process. This leads to inconsistent results in the final part and no way of tracking what input affects what output. The only way to understand this complex process further is to digitally connect each sub-process to the next one, piece by piece and begin to understand the physics of the whole manufacturing process through empirical data.

There are three data sets that are collected: “Print Parameters”, “Melt Pool Metrics”, and “Material Properties”. A “Condition” is a collection of PP’s for a single build, and with that comes a corresponding set of MPM’s and MP’s. By linking together the PP’s, MPM’s, and MP’s, a correlation can be drawn between the input and the output. The ideal Material Property in this project is a Porosity of 0%, or in other words, a fully dense part.

Data Set	Print Parameters	Melt Pool Metrics	Material Properties
Acronym	PP	MPM	MP
Process Step	Before Printing	During Printing	After Printing
Example Variable	Laser Power	Temperature	Porosity

Data Collection

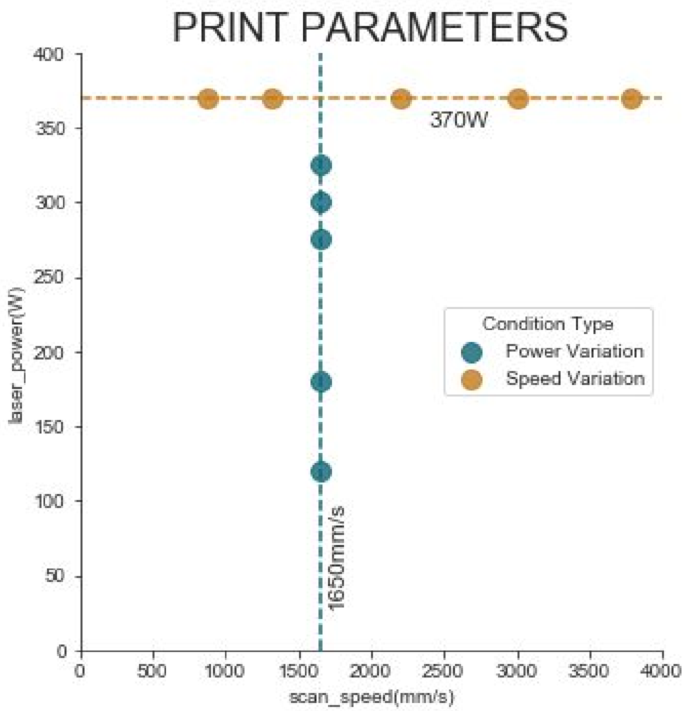
Print Parameters

Print Parameters are set prior to the printing process. These are kept constant throughout the entire build. This project included 10 conditions; 5 with constant scan speed, varying laser power and 5 with constant laser power, varying scan speed.

- **Laser Power** - A diode based laser setting for power in Watts
- **Scan Speed** - The linear speed that the laser moves across the print surface while melting the powder
- **Hatch Spacing** - As the printer is filling in or solidifying the interior of the part, the laser travels in a back and forth motion. Each time it steps over it steps the “hatch spacing distance
- **Condition Type** - Half of the conditions had the same Laser Power while the other half of the conditions had the same Scan Speed
    - PV = Power Variation
    - SV = Speed Variation

- **Layer Height** - Each layer is 0.03mm tall and kept constant throughout the entire project
- **Energy Densities** - A collection of Print Parameters combined into a single variable

Below is a plot of the 10 different conditions with a value for Laser Power and value for Scan Speed. When the Laser Power is constant, it's set to 370W, while when Scan Speed is constant, it's set to 1650mm/s.

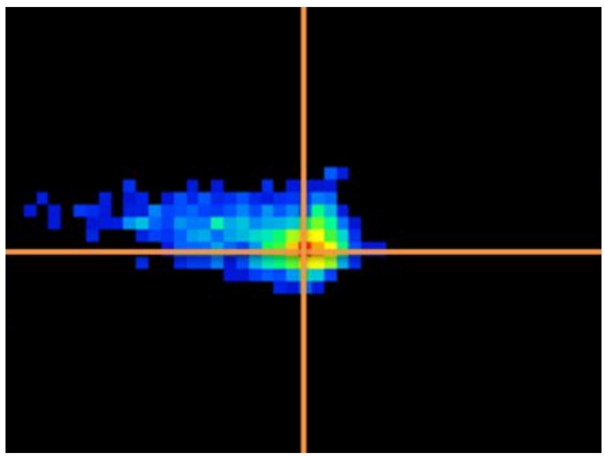


Condition	Laser Power (W)	Scan Speed (mm/s)	Hatch Spacing (mm)	Condition Type	LED	GED	VED
1	325	1650	0.09	PV	0.20	2.19	72.95
2	300	1650	0.09	PV	0.18	2.02	67.34
3	275	1650	0.09	PV	0.17	1.85	61.73
4	180	1650	0.09	PV	0.11	1.21	40.40
5	120	1650	0.09	PV	0.07	0.81	26.94
6	370	3780	0.14	SV	0.10	0.70	23.31
7	370	3000	0.14	SV	0.12	0.88	29.37
8	370	2200	0.14	SV	0.17	1.20	40.04
9	370	1320	0.14	SV	0.28	2.00	66.74
10	370	880	0.14	SV	0.42	3.00	100.11

Each condition has a correlating set of Melt Pool Metrics and Material Properties. Both are described below in the following sections.

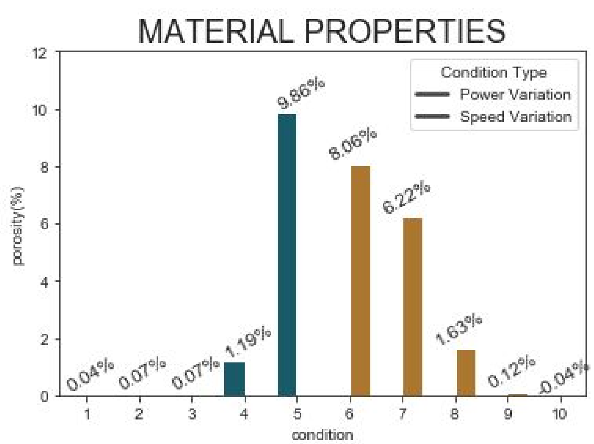
Melt Pool Metrics

The MPM's (x37) were collected through a thermal imaging sensor placed inside the printer. A sample image of a melt pool is shown below; roughly in the shape of a teardrop, traveling to the right. Metrics are calculated and extracted from it and was imported in two seperate files, a viz file and threshold file. The viz file included metrics like intensity, peak pixel location and scan direction. The threshold file included metrics like length, width, area, and temperature of the melt pool. The reason for the second file is because these metrics are based on user defined temperature thresholds of (1600, 1800, 2000, 2200) Celcius. By utilizing a "Runs" table, the viz and threshold tables can be combined into a single table based on which condition they are related to.



Material Properties

The Material Property collected in this project was porosity. A porosity was calculated for each condition, with 0% being the best. A fully dense part increases strength, fatigue life and other mechanical properties for building high value components. Below is a plot of Porosity vs. Condition, ranging from ~0% to ~10%.



Data Wrangling

Print Parameters

The Energy Densities were calculated using the original print parameters set on the machine.

- LED - Linear Laser Energy Density
  - $LED = LP/SS$
  - $LED = Laser\ Power\ /\ Scan\ Speed$
- GED - Global Energy Density
  - $GED = LP/(SS \times HS)$
  - $GED = Laser\ Power\ /\ (Scan\ Speed \times Hatch\ Spacing)$
- VED - Volumetric Laser Energy Density
  - $VED = LP/(SS \times HS \times LH)$
  - $VED = Laser\ Power\ /\ (Scan\ Speed \times Hatch\ Spacing \times Layer\ Height)$

Melt Pool Metrics

Multiple metrics were converted using units, variables, etc. The scan direction was converted from 1 and 2 to x and y. The intensities were normalized using the exposure time. The dimensions length, width, area and peak location were all converted from pixel units to microns using the scale factor. An addition column of Length to Width Ratio was calculated. All the numerical columns were rounded to 2 decimals. The metrics were grouped into 4 categories, general, intensity, peak temperature, and threshold lists. This made is easier to select a subset of the metrics when needed. The last step was to average the huge dataset of metrics per the 10 conditions.

Material Properties

The porosity data was not wrangled and just left as is.

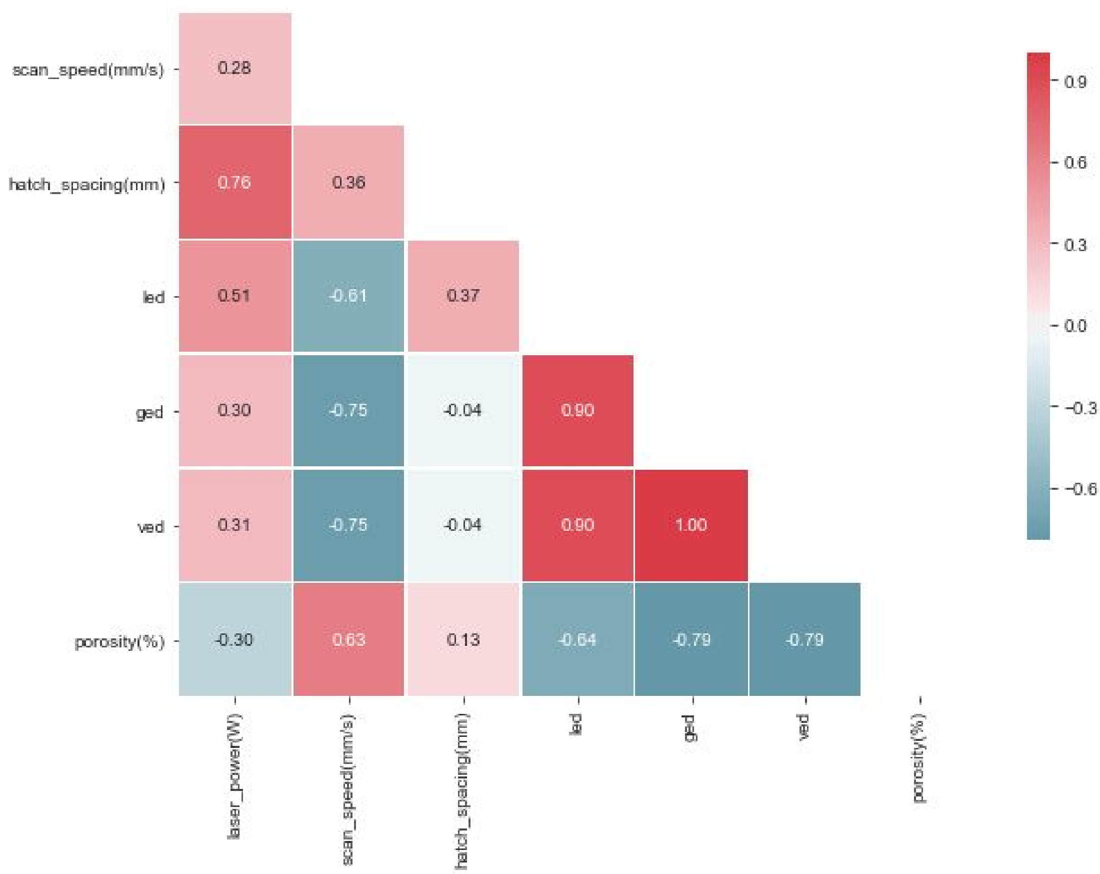
Exploratory Data Analysis

Each of the three datasets were plotted on a correlation plot, similar to the one below. In this example, the Print Parameters were compared to the Material Property: porosity. The scores range from -1 to 1.

- **-1** = Perfectly correlated in opposite directions, as one variable increases, the other decreases
- **0** = No correlation between the data
- **1** = Perfectly correlated in the same direction, as one variable increases, the other increases

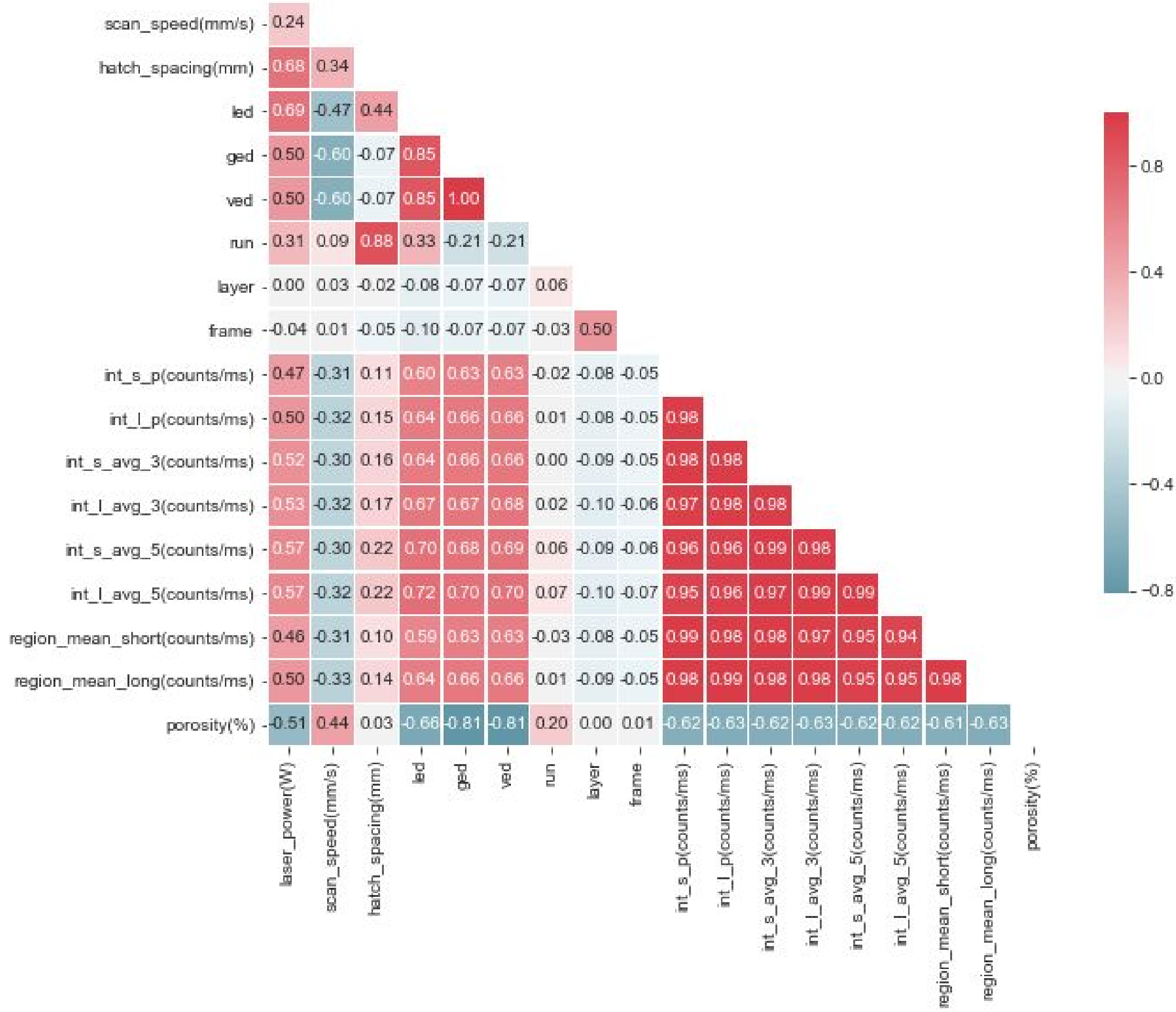
Print Parameters vs. Material Properties

The bottom row of values show the correlation between PP and MP while the rest of the values show the correlation between the individual PP's. GED and VED most directly correlate with Porosity, both values being -0.79. As GED and VED increase, porosity decreases.



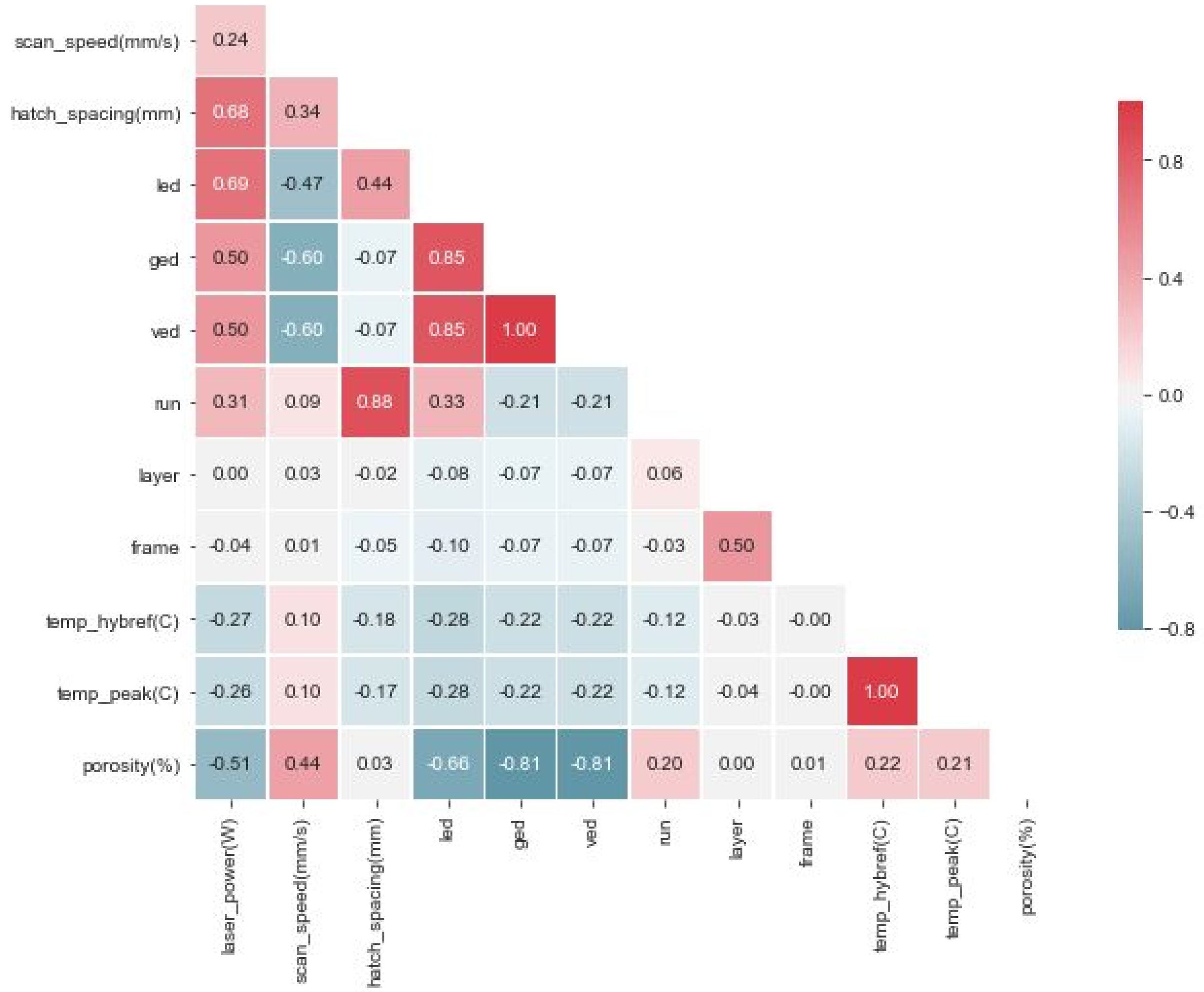
Print Parameters vs. MPM_int vs. Material Properties

The Long Intensity Averaged over 5x5 pixels correlates fairly well with LED, 0.72. All the intensity columns correlate roughly the same with Porosity, between -0.61 and -0.63.



Print Parameters vs. MPM_peak vs. Material Properties

The Hybrid Reference Temperature and Peak Temperature don't really correlate with Porosity or the Print Parameters very well.



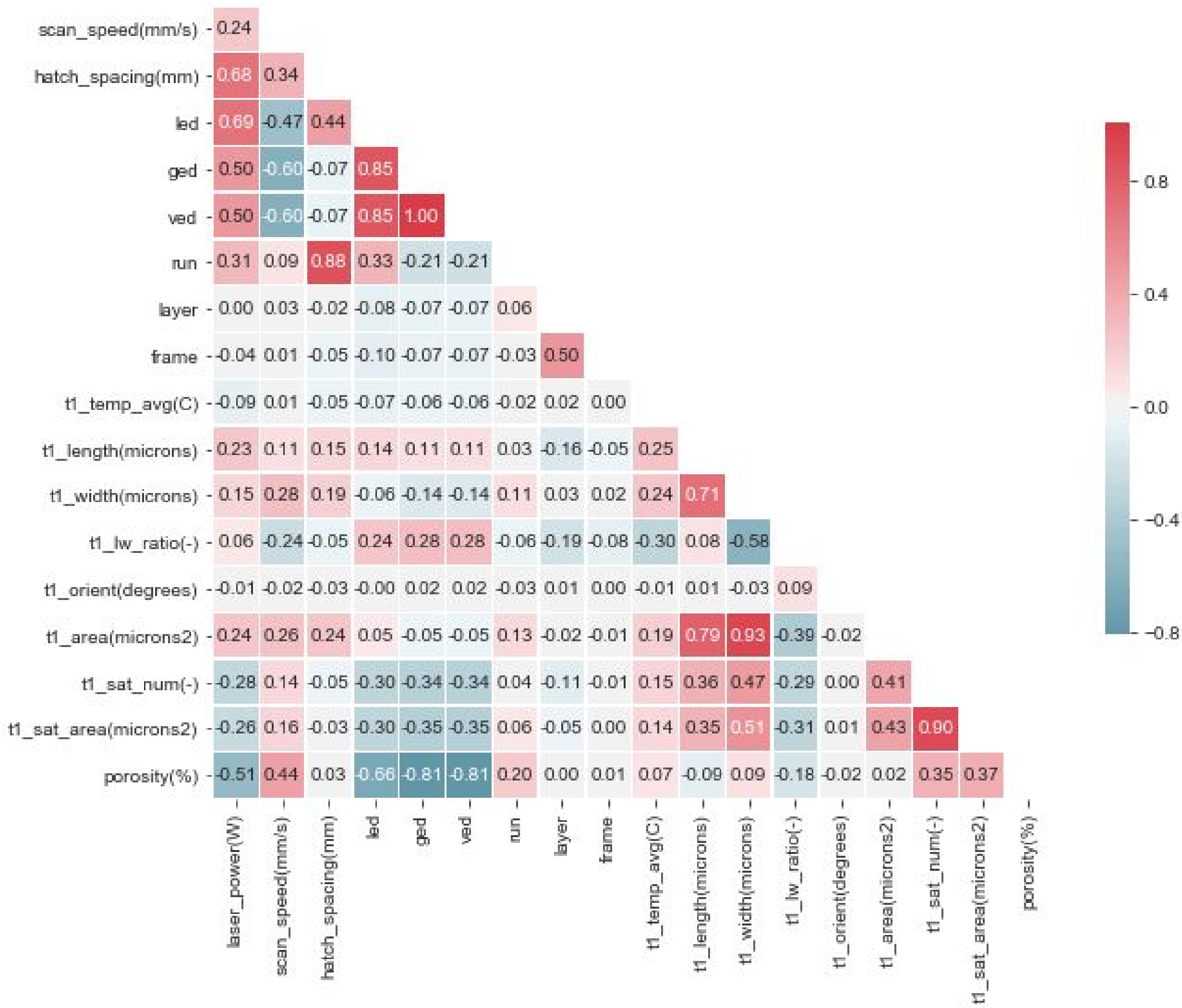
Print Parameters vs. MPM_threshold vs. Material Properties

The Melt Pool Metrics from Threshold 1 (t1) don't correlate well with Porosity, the two highest being the Number of Satellites and the Satellite Area, at 0.35 and 0.37, respectively. Also, they don't correlate very well will the Print Parameters either.

Length, Width, LW_Ratio, Area

- Length vs. Width = 0.71
- Length vs. LW_Ratio = 0.08
- Length vs. Area = 0.79
- Width vs. LW_Ratio = -0.58
- Width vs. Area = 0.93

So Area is affected by both Length and Width, but by Width quite a bit more than Length. (0.93 vs 0.79) For how area is calculated, it makes sense to be highly correlated with Length and Width, but because of the tear dropped shape of the Melt Pool, the Width affects the Area more than the Length. Thresholds 2, 3, and 4 have similar correlations but are generally weaker.



Correlation Conclusion

- PP vs. MP
  - GED/VED are the best indicators for Porosity (-0.79)
- MPM vs. MP
  - Long Peak Intensity Pixel value is the best indicator of Porosity (-0.63)
- PP vs. MPM
  - Long Intensity Averaged (5x5 mask) correlates fairly well with LED (0.72)

Machine Learning Analysis

Steps

1. Data Prep
   a. Rows - Conditions (1-10)
   b. Columns - PP, MPM, MP
   c. Split Data - 70/30
   d. Variance Inflation Factor
2. Evaluation
   a. Root Mean Squared Error
   b. R-Squared
3. Models
   a. Linear Regression
   b. Random Forest

Data Preparation

A couple steps were applied to the data to prepare it for modeling. First, multiple tables were combined into a single dataframe; Print Parameters, Melt Pool Metrics, and Material Properties. This new table had 45 columns and 10 rows, one for each condition. Any column with non-numerical values was converted to numerical values; for example: Condition Type was converted from PV/SV to 0/1. Then the table was split into a Train and Test set (70/30). In this case, that meant that 7 rows were designated for Training, while 3 rows were designated for Testing. From there, 4 additional dataframes were created to later be modeled.

Data Sets to be Modeled

1. All Features
    a. All 44 features are used to predict Porosity
2. VIF Features
    a. 2 Features are used to predict Porosity (based on a Variance Inflation Factor)
3. Print Parameters
    a. All 7 features are used to predict Porosity
4. Melt Pool Metrics
    a. All 37 features are used to predict Porosity

Variance Inflation Factor (VIF)

Collinearity is when two or more variables are highly correlated with one another, in other words, duplicate information within a dataset. Ideally, features in a dataset display different information. Collinearity can cause inflation of the variance of a regression coefficient which may cause predictions with large errors. A VIF value is calculated for each feature/column of data:

- $VIF \approx 1 = No\ Collinearity\ between\ Features$
- $1 \leq VIF \geq 5 = Moderate\ Collinearity$
- $VIF \geq 5 = Serious\ Collinearity$

After removing all columns of data where the VIF is greater than 5, the dataset is left with two columns: the Average Temperature for Threshold 4 and the Area of the Satellite Particles for Threshold 4 [t4_temp_avg (C), t4_sat_area (microns2)].

Evaluating a Model

Root Mean Squared Error (RMSE)

RMSE measures the difference between the values predicted by a model and the values observed. RMSE is always positive and the lower the value, the better the model is at predicting the outcome. Below is the mathematical equation of RMSE:

$$RMSE = \sqrt{\frac{\Sigma_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

R-Squared (R^2)

R^2 is the proportion of the variance in the dependent variable that is predictable from the independent variables. R^2 is always between 0 and 1, and the higher the value, the better the model is at predicting the outcome. Below is the mathematical equation of R^2:

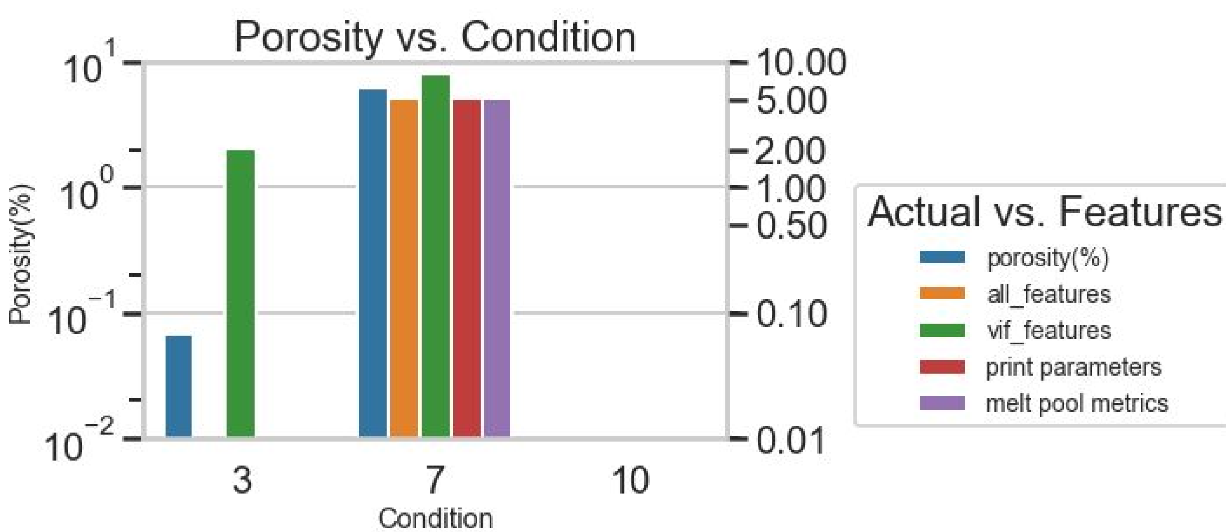
$$R^2 = 1 - \frac{Explained\ Variation}{Total\ Variation}$$

Linear Regression

A function was created to initialize, train and test a Linear Regression model. The larger the coefficient for the feature, the more directly it affects the output, porosity. Below is a table of of the top 3 features for each model with the 4 different datasets. The Print Parameters have very large coefficients compared to the other datasets, with GED being the largest at 1400.

All Features		VIF Features		Print Parameters		Melt Pool Metrics	
Satellite Area T2	2.51e-4	**Average Temp T4**	1.98e-2	**GED**	1400	**Satellite Area T2**	2.51e-4
Short Intensity AVG-5	1.24e-4	**Satellite Area T4**	3.88e-4	**LED**	-304	**Short Intensity AVG-5**	1.24e-4
Satellite Area T3	1.12e-4			**VED**	-41	**Satellite Area T3**	1.12e-4

Below is a plot of the Porosity predictions for the test data (conditions 3, 7, 10) on a y-log scale.



Below is a table of the RMSE and R-Squared values for each of the data sets. Ideally the best model utilizes the least number of features, has the lowest RMSE value, and the highest R-Squared value. With that said, the Linear Regression Model utilizing the Print Parameters dataset creates the best model.

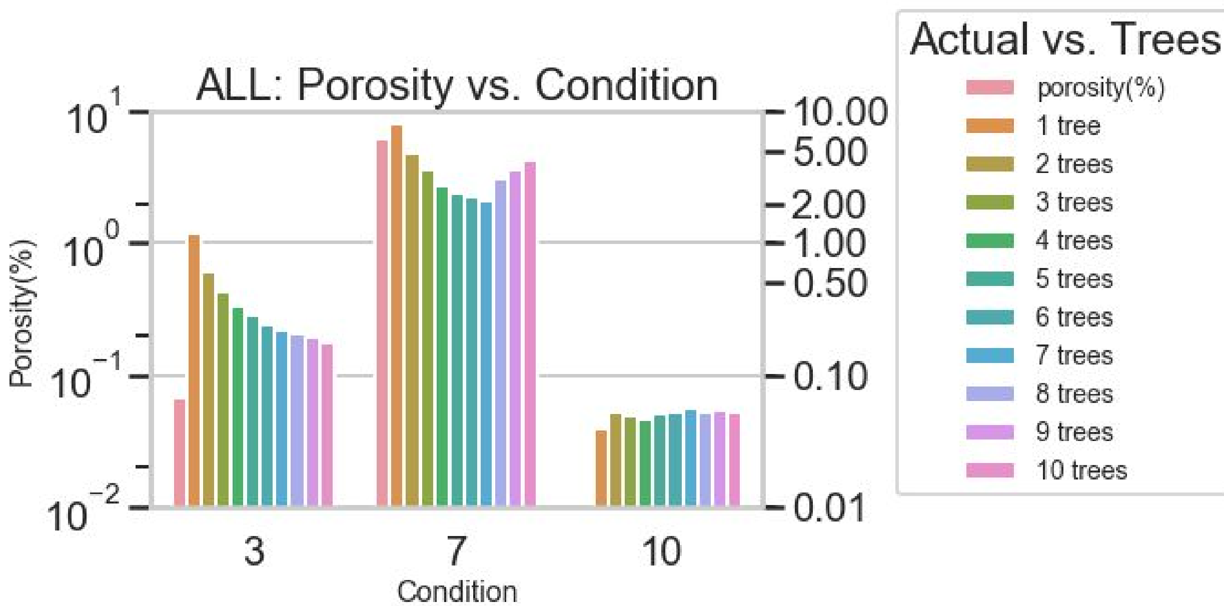
Dataset	# of Features	RMSE	R-Squared
All Features	44	0.641	0.952
VIF Features	2	1.485	0.741
Print Parameters	7	0.638	0.952
Melt Pool Metrics	37	0.638	0.952

Random Forest

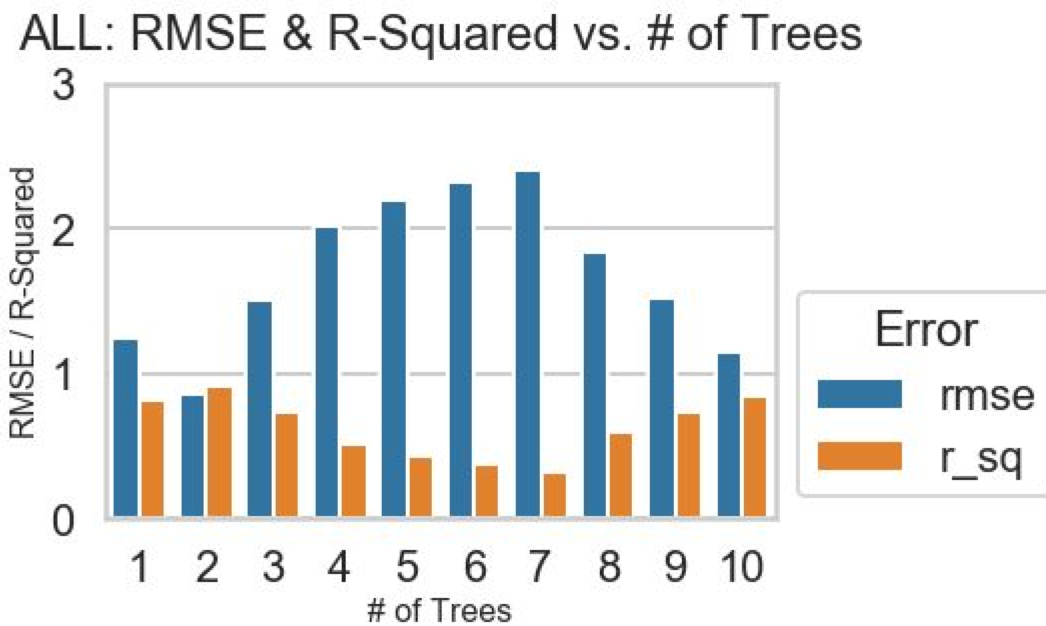
Two functions were created for the Random Forest models. The inner function initializes, trains, and tests the model. The out function creates a for loop of the inner function with a varying number of trees for the model, in this case ranging from 1-10.

All Features

Below is a plot of the Porosity predictions for the test data (conditions 3, 7, 10) on a y-log scale.



Below is a plot of the RMSE and R-Squared values for each of the 10 models created with a varying number of trees. The model with the lowest RMSE and the highest R-Squared values used 2 trees.

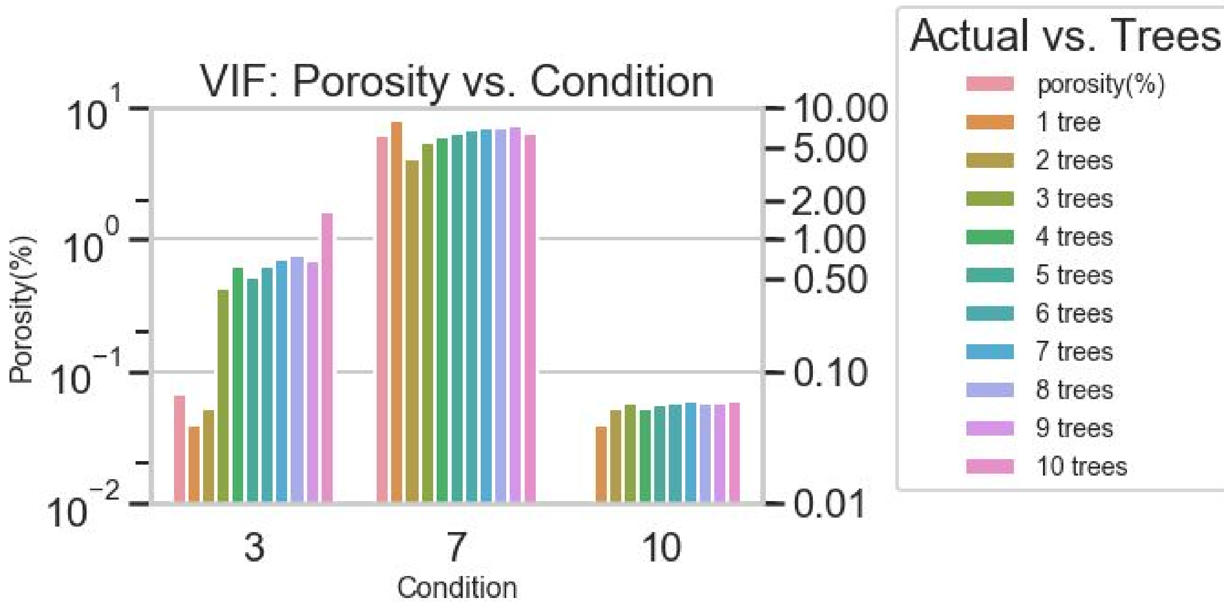


Below is a table representing the three most important features in the model. This model had a RMSE value of 0.858 and an R-Squared value of 0.914.

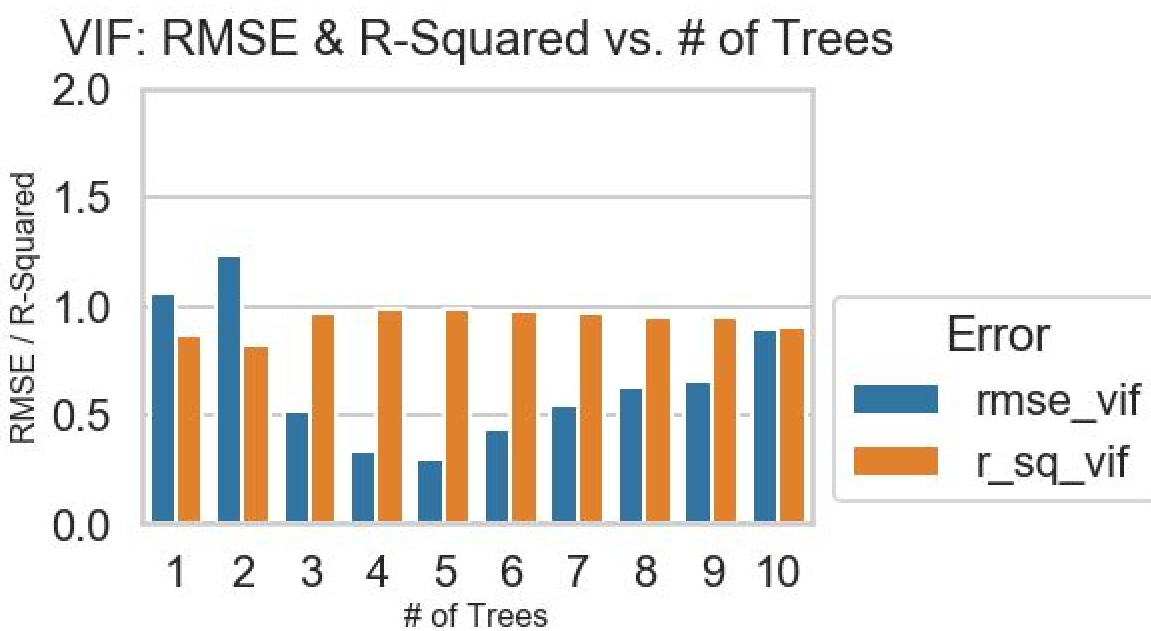
Feature	Importance
region_mean_long (counts/ms)	0.499
t4_sat_num (-)	0.477
t1_sat_area (microns2)	0.020

VIF Features

Below is a plot of the Porosity predictions for the test data (conditions 3, 7, 10) on a y-log scale.



Below is a plot of the RMSE and R-Squared values for each of the 10 models created with a varying number of trees. The model with the lowest RMSE and the highest R-Squared values used 5 trees.

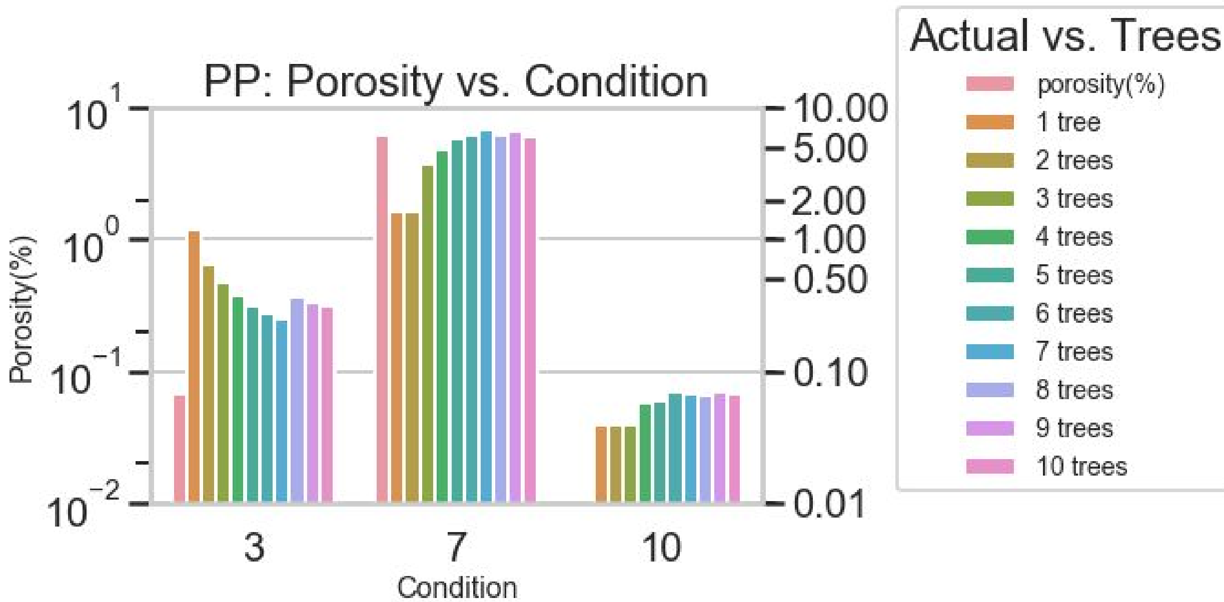


Below is a table representing the three most important features in the model. This model had a RMSE value of 0.295 and an R-Squared value of 0.990.

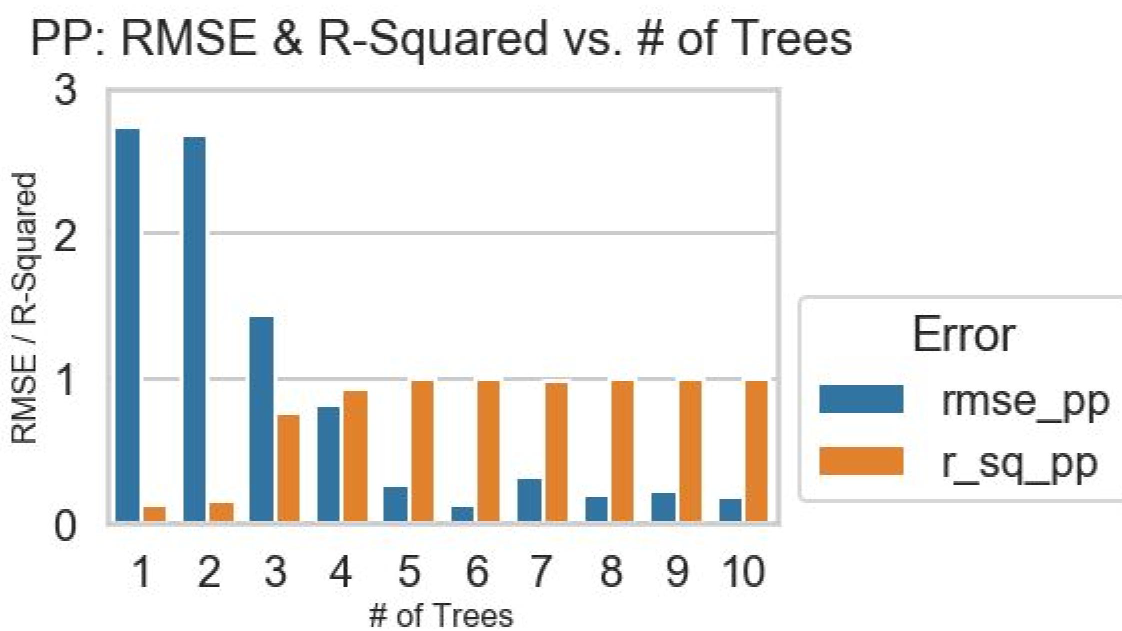
Feature	Importance
t4_sat_area (microns2)	0.788
t4_temp_avg (C)	0.212

Print Parameters

Below is a plot of the Porosity predictions for the test data (conditions 3, 7, 10) on a y-log scale.



Below is a plot of the RMSE and R-Squared values for each of the 10 models created with a varying number of trees. The model with the lowest RMSE and the highest R-Squared values used 6 trees.

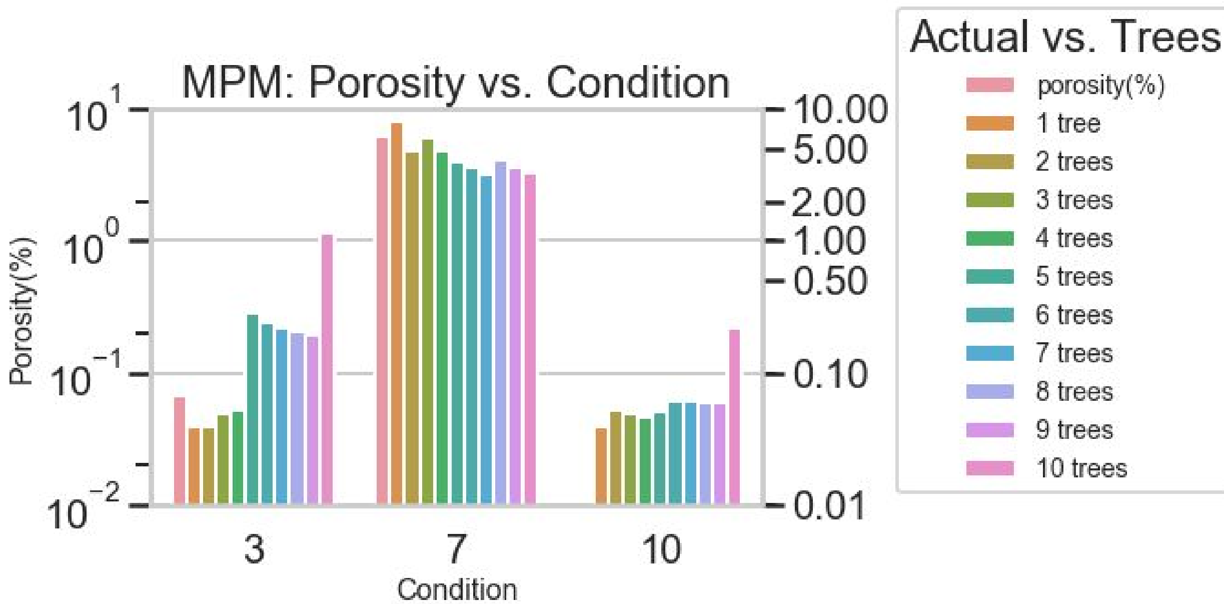


Below is a table representing the three most important features in the model. This model had a RMSE value of 0.124 and an R-Squared value of 0.998.

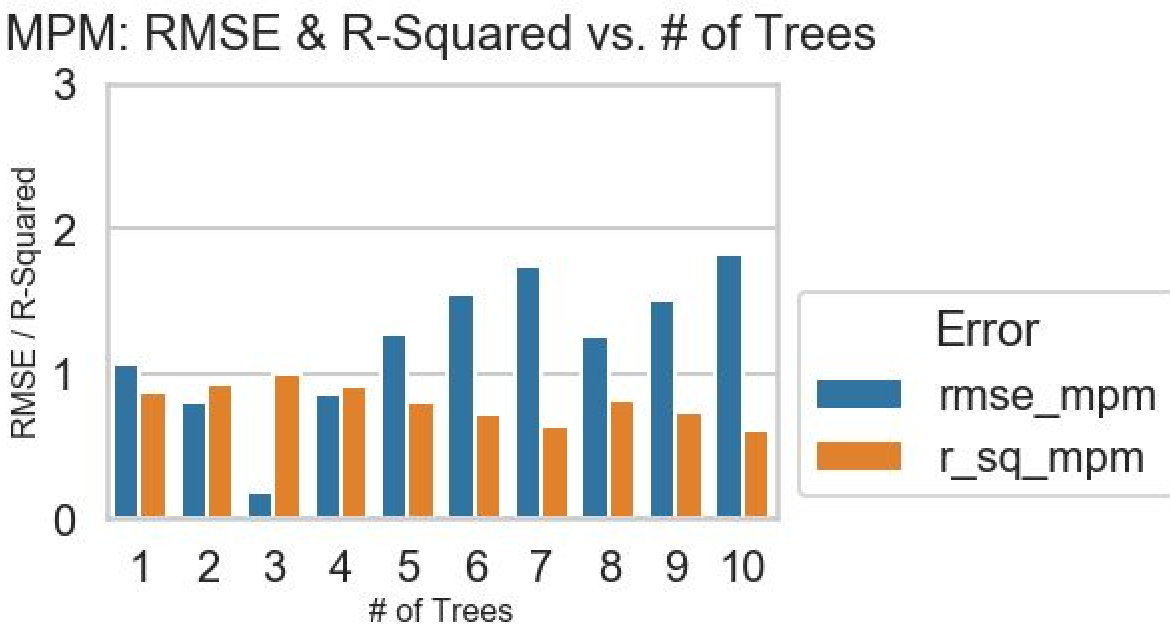
Feature	Importance
VED	0.494
Scan Speed (mm/s)	0.323
LED	0.173

Melt Pool Metrics

Below is a plot of the Porosity predictions for the test data (conditions 3, 7, 10) on a y-log scale.



Below is a plot of the RMSE and R-Squared values for each of the 10 models created with a varying number of trees. The model with the lowest RMSE and the highest R-Squared values used 3 trees.



Below is a table representing the four most important features in the model. This model had a RMSE value of 0.180 and an R-Squared value of 0.996.

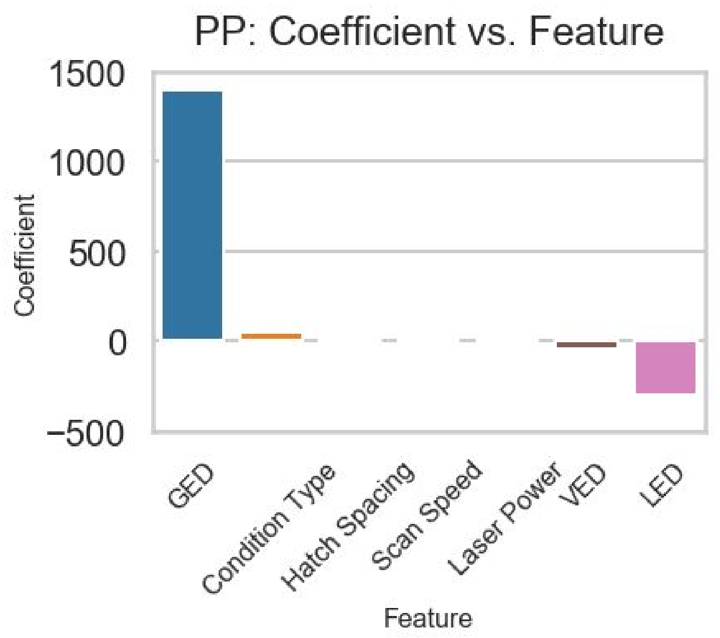
Feature	Importance
t2_sat_num (-)	0.641
int_s_p (counts/ms)	0.333
t4_sat_area (microns2)	0.013
t2_sat_area (microns2)	0.011

Best Models

Linear Regression

The best Linear Regression model is with the Print Parameters as input features. Below is a table of the coefficients in the equation per each feature. This model utilized all 7 features and received a RMSE value of 0.638 and a R-Squared value of 0.952.

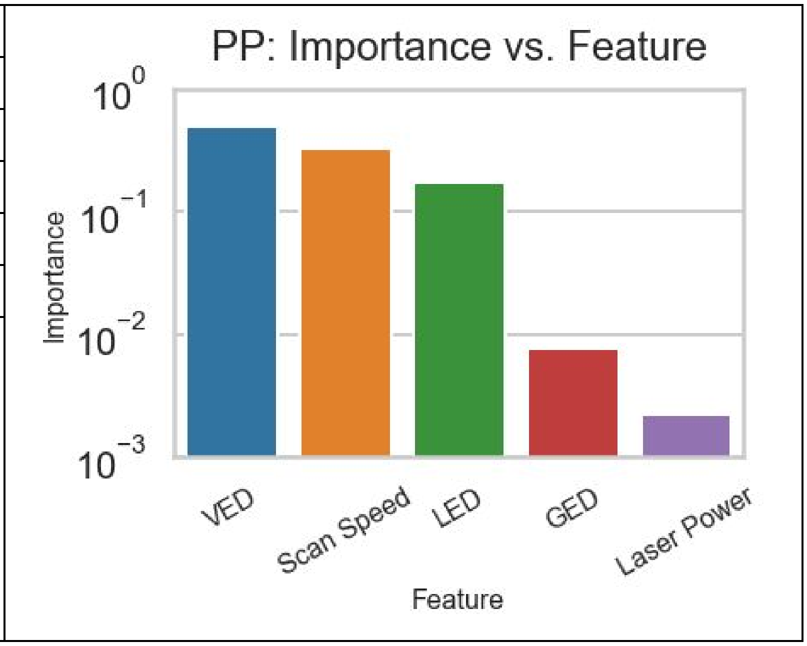
Feature	Coefficient
GED	1400
Condition Type	46
Hatch Spacing	2.30
Scan Speed	0.00
Laser Power	-0.15
VED	-41
LED	-304



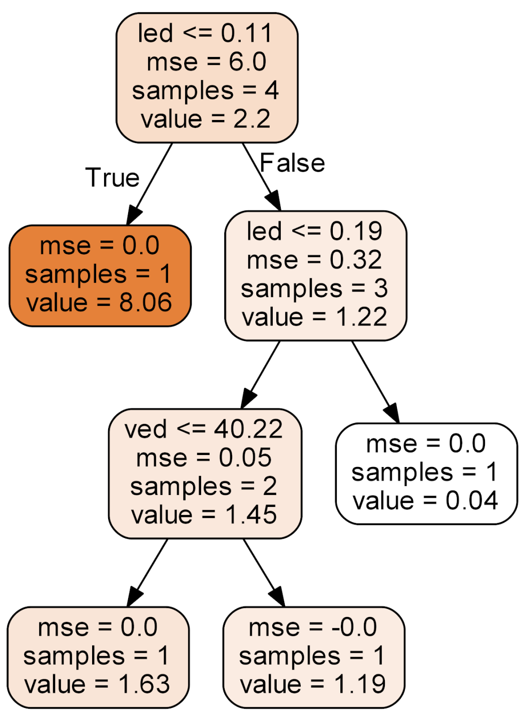
Random Forest

The best Random Forest model is with the Print Parameters as input features. Below is a table of the importance of each feature. This model utilized all 7 features and received a RMSE value of 0.124 and an R-Squared value of 0.998.

Feature	Importance
VED	0.494
Scan Speed	0.323
LED	0.173
GED	0.008
Laser Power	0.002



This is an example of a tree in the Random Forest model.



Conclusion

The best way to increase yield is to reduce porosity in the printed parts. The user's most direct method of achieving this is by controlling the VED, or Volumetric Laser Energy Density, which is a combination of Print Parameters.

Looking Forward

Increase the Number of Conditions

Currently, 10 conditions were selected. 5 with the same Laser Power, 5 with the same Scan Speed. It may be beneficial to select other pairs of Print Parameters to further understand the printing process and how it's affected by the user inputs.

Image Analysis

Currently, Melt Pool Metrics were extracted from the images to create numerical datasets for further analysis. It may be beneficial to use image analysis on the melt pool images directly to find patterns that may not be understood through the metrics.

Higher Resolution of Porosity

Currently, Porosity was measured over the entire volume of each condition. This lead to averaging the Melt Pool Metrics over the entire build. It may be beneficial to increase the resolution of the porosity value for each condition. This may increase the correlation between the two data sets.