Project 4

AGENDA

- Team introductions
- Project Background
- Data models
- Conclusions

Meet Team 4

Mica Zier

Bobbi Fletchall

Linda Jepkorir

Project Background & Concept

Wonder Woman Construction Company Inc (WWC) currently bids for jobs based on an industry software program that calculates job times via tasks. The task times used by the software are set based on historical company data (assumed to be outdated by company leadership), and WWC Inc. would like to build a tool that uses their current company task time benchmarks to estimate the bids.

WWC has not attempted this task before because they did not have a robust data set of current historical time spent per job. A new data collection tool was recently put in place at WWC, and this new data will be used to build a new model to estimate Job Times. If the new model build is successful, WWC could then estimate bids base on their current company data versus outdated historical data.

The key questions this Analysis will answer are:

- Is the new data collection tool able to provide enough current historical data for a new model to be built?
- What are the most important factors for estimating job time, and is the new data collection tool able to provide it?
- Can a new model with a minimum of 80% accuracy be created from the current historical data?
- How should WWC Inc. change their bid strategy based on this data?

WWC data was provided for analysis, which included Estimated Time and Actual time for each job. Team reviewed data file to determine meaning and significance of each column for project

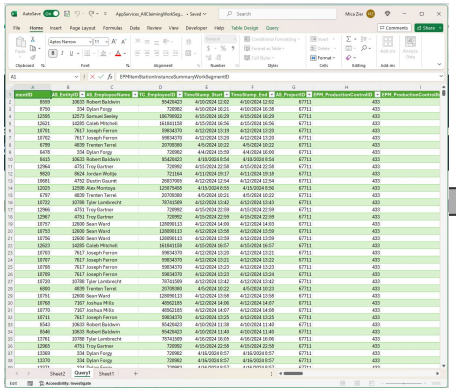
Field	Definition	Relevant to Project?
Assembly_MainPieceProductionCode	Piece Category: Beam, Pipe, Door Plate, Backer Bar, Crane Grider, Bollard, Alum Pipe, Galv Stringer, etc.	Yes
Assembly_MainPartDimension	Inches by pound	Yes
Assembly_MainPartFinishDescr	Abbreviation + finish of the part description	Yes
Assembly_MainPartGrade	Grade of instance	Yes
Assembly_MainPartLengthFt	Part length	Yes
Assembly_MainPartShape	Shape of instance	Yes
Assembly_NumSmallParts	Number of things to assemble per instance (not per employee per instance)	Yes
Assembly_SurfaceAreaEachSqFt	Part surface area	Yes
Assembly_TotalQuantityInJob	Quantity per instance	Yes
Assembly_WeightEachLbs	Part weight	Yes
EPM_AdjustedStationName	Use this column to identify which task of the instance it is on. One instance can contain multiple station names (or tasks).	Yes
EPM_ProductionControlItemID	Unique identifier of a group of instances. Unique to job number and main mark.	Yes
EPM_InstanceNumber	Instance Number - we want to find time spent per instance	Yes
TimeInSeconds_ThisWorkSegment	Time spent per instance in seconds - no need to calculate time from start and end, it done already	Yes, this is unit of time it takes per instance. TARGET OF MODEL
Assembly_EstTotalHours_ThisLaborGroup	Software Estimated time, in hours. This is what we are comparing to.	Yes, as this is current estimate of time being used by WWC to bid jobs, but needs to be converted to seconds.

Field	Definition	Relevant to Project?
EPM_MainMark	Indentifer of group of ControlItemIds, which then have a group of instances	Yes, but is a descriptor that is not necessary to model
NumberWithDash	Project number. Unique to site/project. Will contain multiple Main Marks that have mutiple ControlItemids which then have mutiple instances	Yes, but is a descriptor that is not necessary to model
AdjustedEnd	Not in use, so remove	No
AdjustedStart	Not in use, so remove	No
AS_EntityID	Remove	No
AS_ProjectID	AppServices Project ID - used in the software. Unique to each project. But use Project Number instead, this is not needed.	No
Assembly_MainPartFinishAbbrev	Finish of the part abbreviated - do not use	No
EPM_LaborGroupID	Not needed	No
EPM_LaborGroupName	Do not use for model, use later	No
EPM_LaborGroupNumber	Do not use for model, use later	No
EPM_LotNumber	Not relevant	No
EPM_PhaseNumber	Not relevant	No
EPM_ProductionControlID	Software Name Production Control ID. Not needed	No
EPM_SequenceID	Lot sequence - break job down into parts, not relevant	No
EPM_StationID	Not needed	No
EPM_StationName	Remove	No
EPMItemStationInstanceSummaryWorkSegmentID	Remove	No
JobName	Name of the site or project they are working on	No

Field	Definition	Relevant to Project?
STS_Barcode	Not needed	No
TimeClcok_WorkSegmentEnd	Remove	No
TimeClock_WorkSegmentStart	remove	No
TimeClockWorkSegmentRecordID	Remove	No
TimeInSeconds_AdjustedTime	Not in use, so remove	No
TimeInSeconds_TimeClockWorkSegment	Don't need	No
TC_EmployeeID	Time clock worker id	No - no PII
AS_EmployeeName	Name of the worker	No - no PII
TimeStamp_End	End of the instance work	No, calc'd in other column
TimeStamp_Start	Start of the instance work	No, calc'd in other column

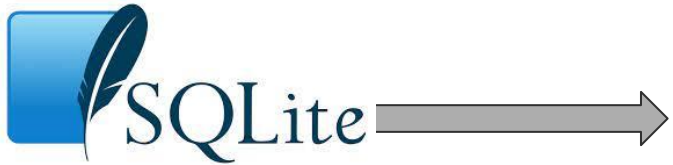
Data model was developed using SQLite, which delivered data to Jupyter notebooks to run the model analysis

WWC Job Data Cleaned & prepped



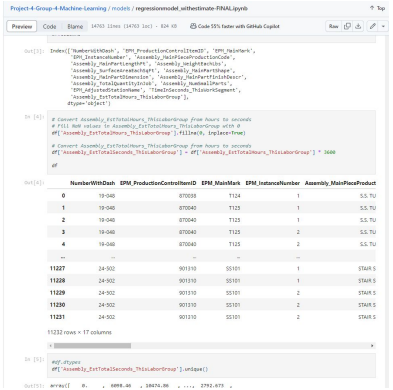
- WWC data was exported out of company tools via .csv file
- Data prep steps

Created Database



- Loaded data into SQLite database, to be read into Jupyter Notebooks

Jupyter Notebooks



- Model Design & Optimizations

Model Review

Identified Features & Target for Model, and removed unnecessary fields

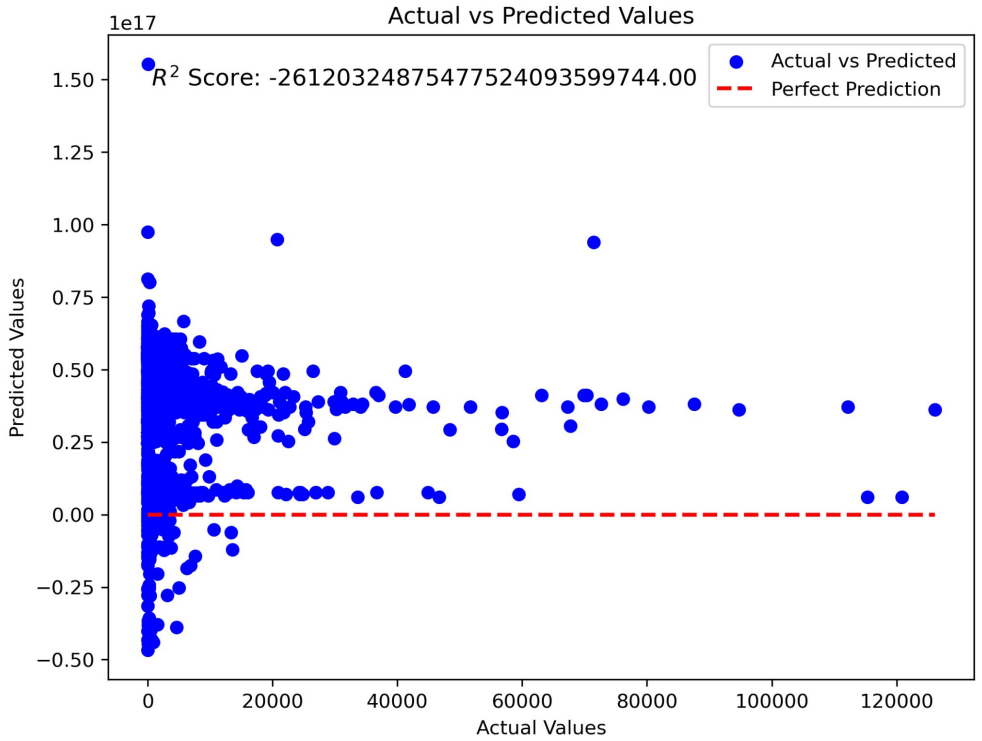
Column Included in Model	Target or Feature?
EPM_ProductionControlItemID	Feature
EPM_InstanceNumber	Feature
Assembly_MainPieceProductionCode	Feature
Assembly_MainPartLengthFt	Feature
Assembly_WeightEachLbs	Feature
Assembly_SurfaceAreaEachSqFt	Feature
Assembly_MainPartShape	Feature
Assembly_MainPartDimension	Feature
Assembly_MainPartFinishDescr	Feature
Assembly_TotalQuantityInJob	Feature
Assembly_NumSmallParts	Feature
EPM_AdjustedStationName	Feature
TimeInSeconds_ThisWorkSegment	Target

Column to Remove	Reason to remove
EPM_MainMark	Descriptor of Job, not impactful variable
NumberWithDash	Descriptor of Job, not impactful variable
Assembly_EstTotalHours_ThisLaborGroup	Outcome variable we'll compare against result of new model

Linear Regression Model was executed, but data was found to have no linear relationship

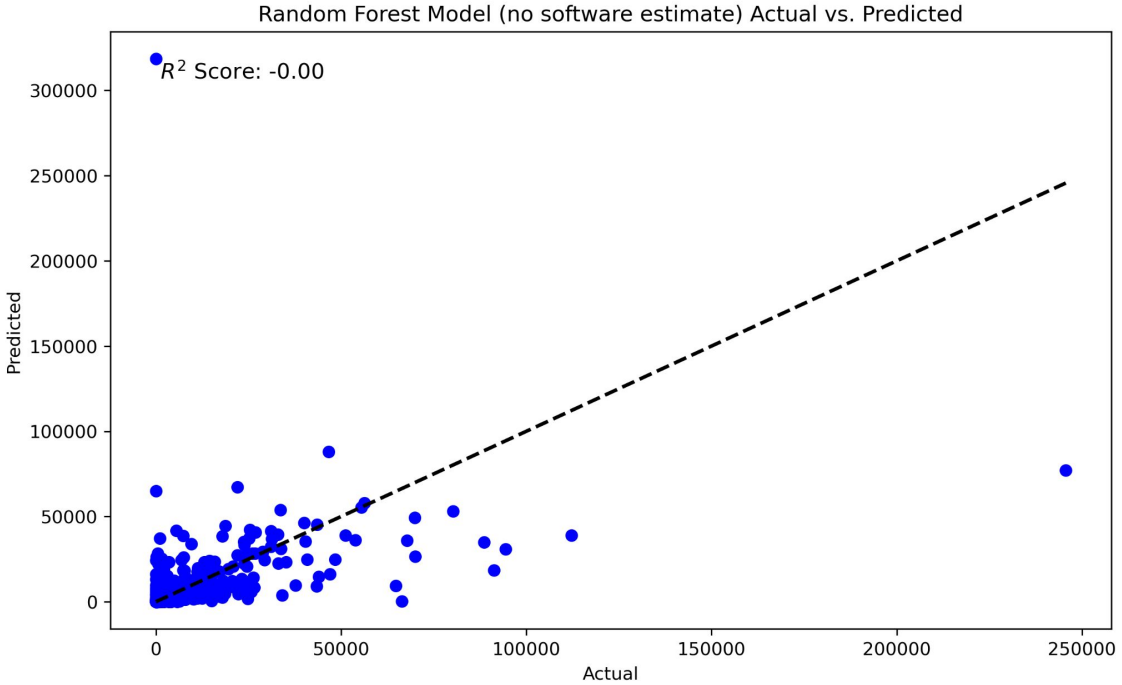
Model 1 - Linear Regressions

- Started with this model to determine correlation of variables



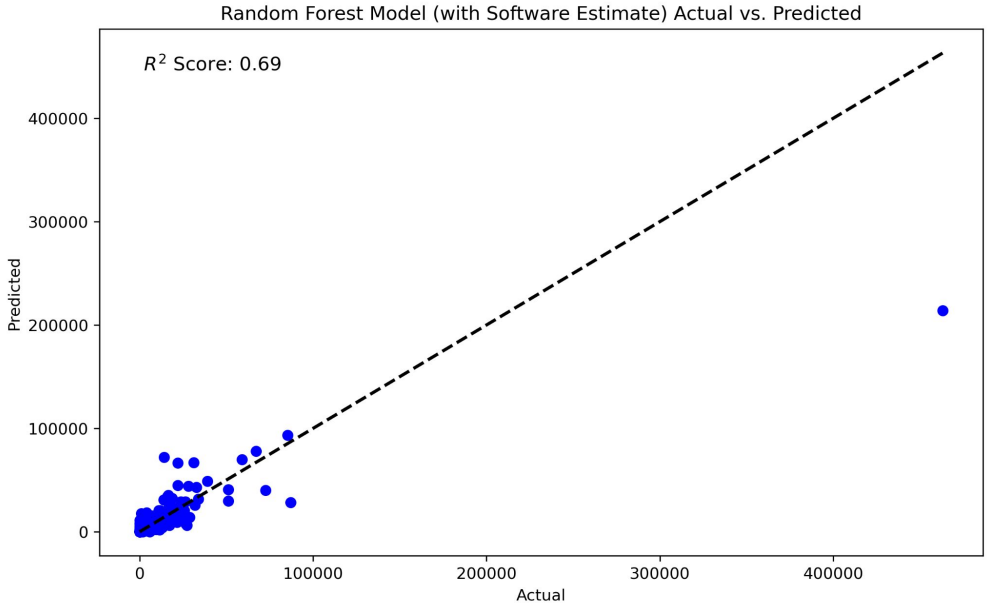
Random Forest Regression Model was tried next as it can capture non-linear relationship between variables

- Optimizations were also applied to the model including:
  - Binning of "Assembly_MainPieceProductionCode" to only include the 50 most commonly used variables, as this would simplify the complexity of the model and hopefully identify a pattern to predict time per Job.
  -
- However, optimization and design of Random Forest Model did not result in 80% accuracy.



A revised Random Forest Model was executed again, but this time the Software Estimate time was included in model to determine if it would help improve accuracy.

- Software Estimate is estimate based on historical company data—although company leadership thought it was outdated
- Converted "Assembly_EstTotalHours_ThisLaborGroup" (the “Software Estimate”) to seconds to ensure it aligned with actual job time field "TimeInSeconds_ThisWorkSegment".
- Included same binning strategy from first Random Forest Model design
- Results drastically improved, with the result improving to 69% accuracy

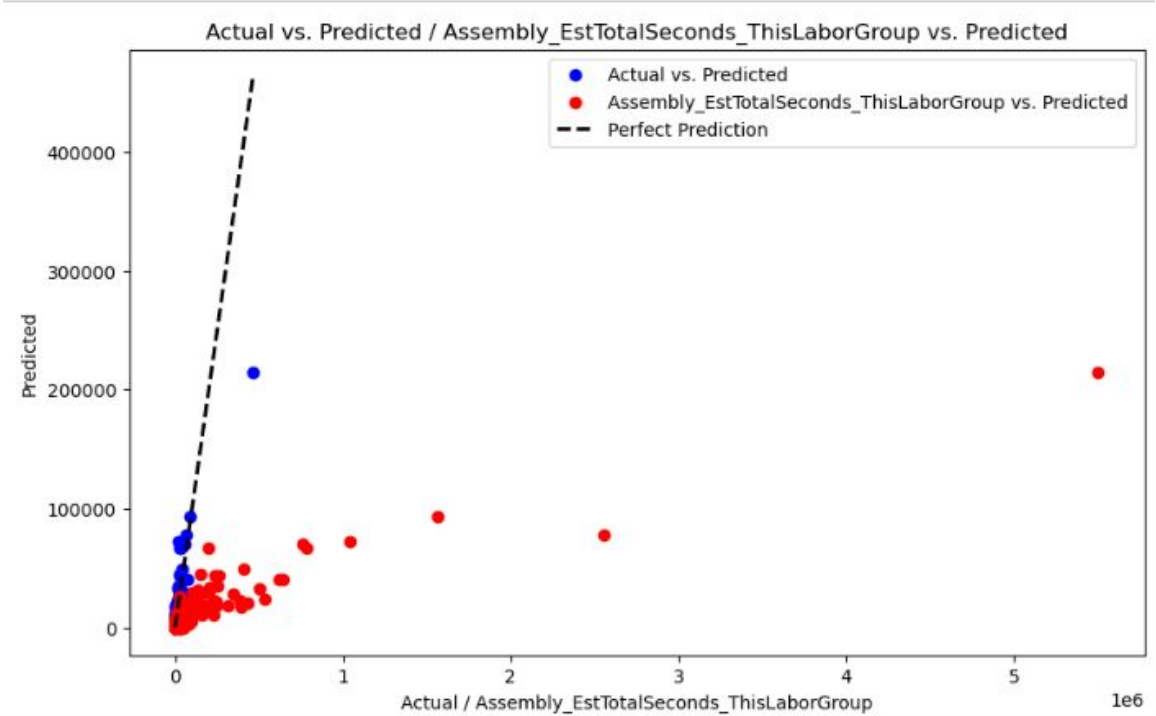


Final Model Selection

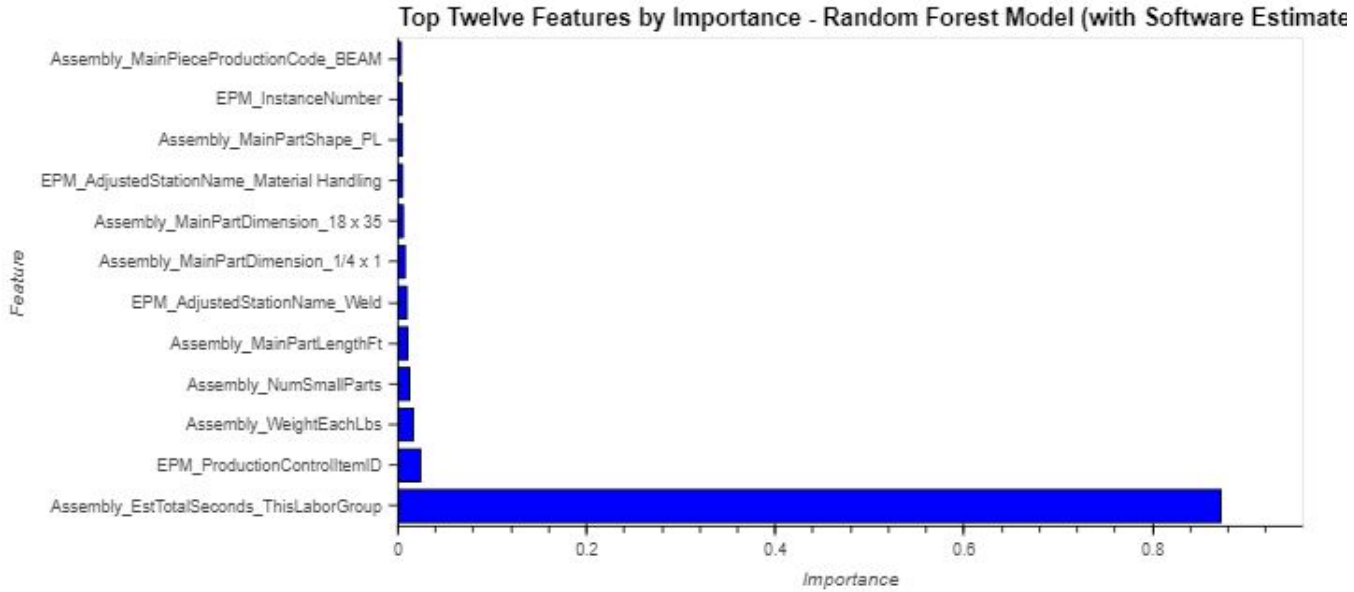
	Model 1: Linear Regression	Model 2: Random Forest without Software Estimate	Model 3: Random Forecast with Software Estimate
Features	ProductionControlItemID InstanceNumber MainPieceProductionCode MainPartLengthFt WeightEachLbs SurfaceAreaEachSqFt MainPartShape MainPartDimension MainPartFinishDescr TotalQuantityInJob NumSmallParts AdjustedStationName	ProductionControlItemID InstanceNumber MainPieceProductionCode MainPartLengthFt WeightEachLbs SurfaceAreaEachSqFt MainPartShape MainPartDimension MainPartFinishDescr TotalQuantityInJob NumSmallParts AdjustedStationName	ProductionControlItemID InstanceNumber MainPieceProductionCode MainPartLengthFt WeightEachLbs SurfaceAreaEachSqFt MainPartShape MainPartDimension MainPartFinishDescr TotalQuantityInJob NumSmallParts AdjustedStationName EstTotalSeconds (Software Estimate of Time)
Target	TimeInSeconds (Actual)	TimeInSeconds (Actual)	TimeInSeconds (Actual)
Model Evaluation	Linear Regression R^2 Score: -2.612 Linear Regression Mean Absolute Error: 3.853	Random Forest Regression Score (R^2): -0.002 Random Forest Regression MAE: 1465.566	Random Forest Regression Score (R^2): 0.691 Random Forest Regression MAE: 1854.615
Model Selection	Not recommended	Not recommended	Best option with highest accuracy score

Model Comparisons

New Model does not improve upon current estimates provided by WWC Software



Review of Feature Importance indicate Software Estimate is the top contributor to predicting job time



Conclusions

Conclusions

- Project did not result in model with minimum of 80% accuracy
- WWC should continue to use Software Estimator Tool while the new data collection process is put in place, as the current most important factor in predicting time for jobs is the Software Estimate
- Once additional data is collected, ensure it reflects only completed jobs to make sure time reflect completed jobs v. jobs still in progress
- Try the model build exercise again with updated data

Q & A