Predicting the cost of driver pay to increase the accuracy of future bids

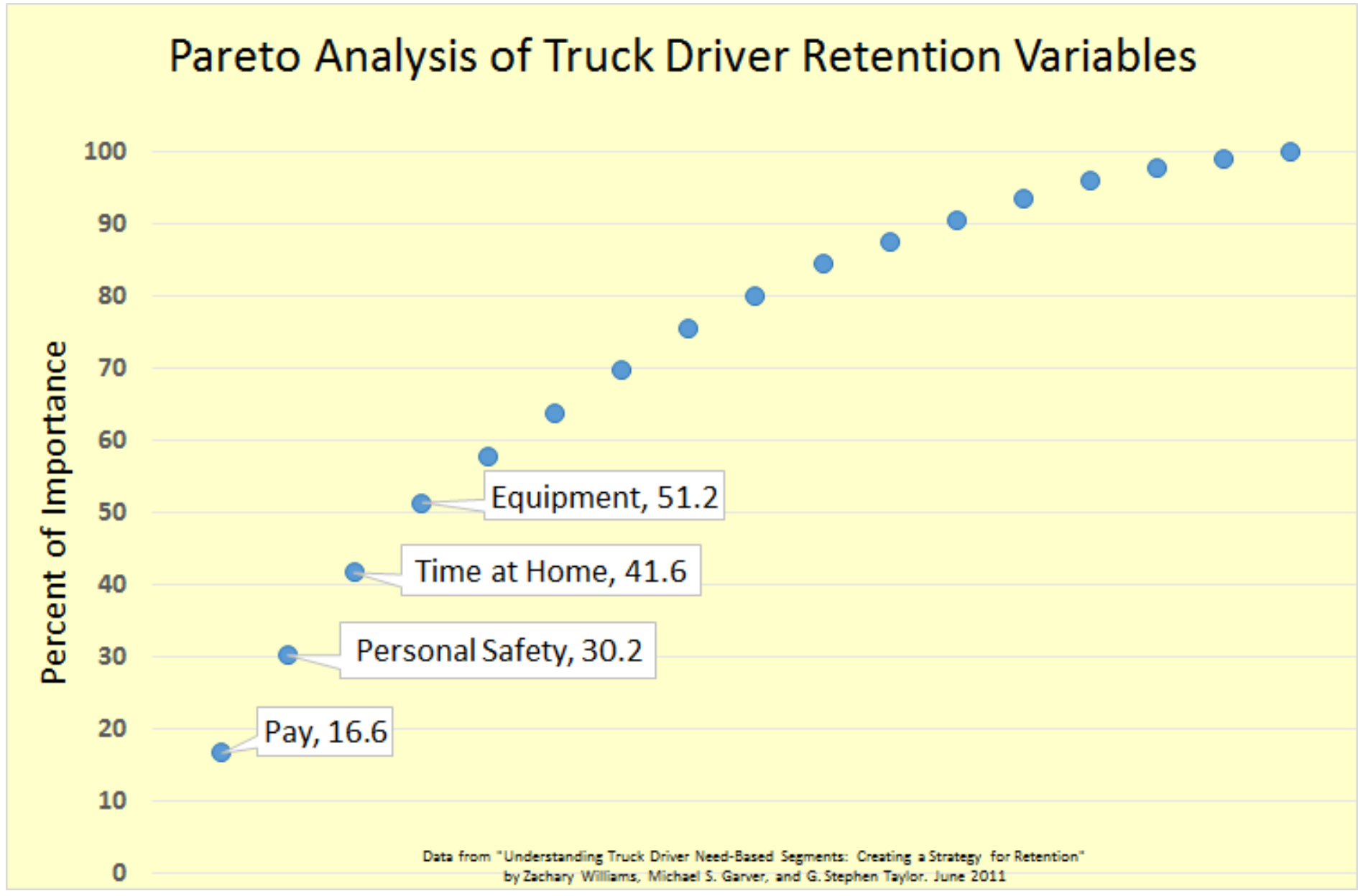
Caine McLeod, Foster Pollock, Jackson Goodman, Zachary Osiecki

Abstract

The trucking industry, as a whole, has a very high turnover rate for drivers. The turnover rate can vary wildly from 40% to as high as 200% on an annualized basis. When bidding a contract for potential customers, estimating the cost of driver pay can be very difficult because of the volatility in the market. Currently, J.B. Hunt relies heavily on industry experts to estimate these costs.

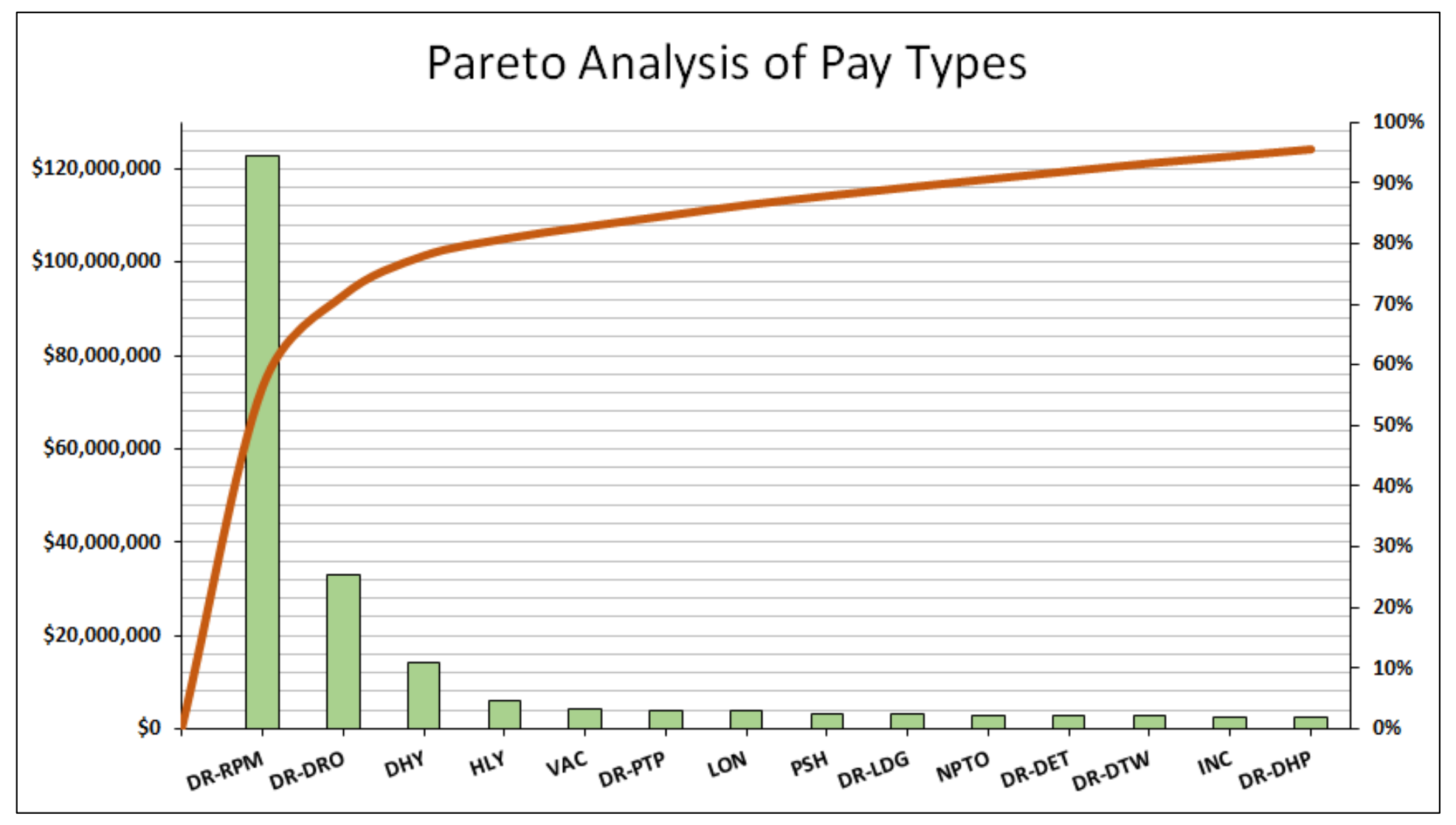
Our team decided to create a decision support tool to help ground these estimates in two sets of data. The first set, J.B. Hunt's internal data, was analyzed using many different data analysis techniques to determine an appropriate regression model for predicting driver pay. The second set, from the Bureau of Labor Statistics, is automatically collected and used to validate each prediction. We used VBA in Excel to collect and gather all the data necessary for the data analysis to be done automatically. The program was designed to be easy to interpret for future development and flexible enough to allow for new datasets to be added without needing to re-write the program. Finally, the end-user is presented with a summary of the analyses performed, so he or she is better equipped to make an estimate.

Preliminary Analysis


Pareto Analysis of Truck Driver Retention Variables

Data from "Understanding Truck Driver Need-Based Segments: Creating a Strategy for Retention" by Zachary Williams, Michael S. Garver, and G. Stephen Taylor. June 2011

Academic research suggests that truck driver retention can be affected by many variables but driver pay was the most influential (Williams, Garver, and Taylor 2011). With shifting demographics and an aging workforce, truck driving is not a desirable occupation for young people. One common complaint from drivers is the high variance of weekly pay, which presented itself in the data.

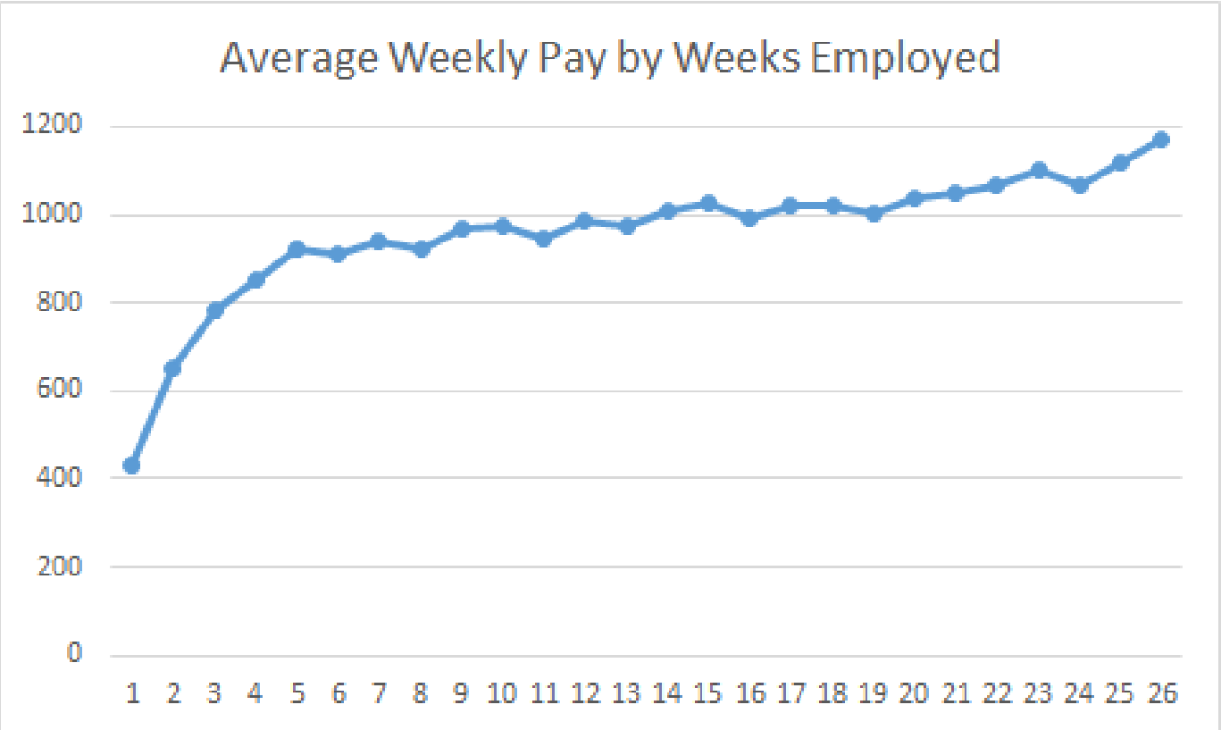

Pareto Analysis of Pay Types

Earning Code	Description
DR-RPM	Driver-Rate Per Mile
DR-DRO	Driver-Drops
DHY	Driver-DCS Hourly Rate
HLY	Hourly Regular
VAC	Vacation Pay
DR-PTP	Driver-Per Trip Pay
LON	Driver-Loaned Non Mileage Pay
PSH	Personal, Sick, Holiday
DR-LDG	Driver Loading
NPTO	Paid Time Off
DR-DET	Driver-Detention
DR-DTW	Driver-Training Wages
INC	Bonus-Incentive Pay
DR-DHP	Driver-Drop and Hook

Pareto analysis for 26 weeks of data suggest 2.8% of earning codes (miles and stops) account for over 70% of driver pay across all accounts. Driver Rate-Per-Mile and a calculated average stops-per-week were used as continuous for all candidate models.

Modeling

Data was examined from two different perspectives, at the individual driver level and at the account level, taking averages for miles and stops across the number of drivers for the given week. Multiple Linear Regression was the preferred modeling technique used in Minitab 18. Inaccurate unit amounts for mileage pay, weekly pay variance due to nature of activity-based pay, and increasing average weekly pay by number of weeks worked led us to examine driver pay from the account level.


Average Weekly Pay by Weeks Employed

Assumptions

Weekly account data was analyzed and separated manually through the use of pivot tables in excel. Rows were tabularized with Account number, driver class, region, channel, industry, and pay end date as the unique key. Through backwards elimination, binary variables categorizing activities performed by account drivers were eliminated based on a model fit criterion of alpha = 0.01. Binary variables with negative coefficients were excluded from the final model, as they did not reflect an increase in pay for added driver activity.

Model

Variables

$X_1 = Average\ Weekly\ Miles, X_2 = Average\ Weekly\ Stops, X_3 = Job\ Family\ Class$

$X_4 = Customer\ Region, X_5 = Customer\ Channel, X_6 = Customer\ Industry$

$X_7 = Estimated\ Difficulty, X_8 = Live\ Load\ (0{,}1), X_9 = Drop\ \&\ Hook\ (0{,}1),$

$X_{10} = Tarping\ (0{,}1)$

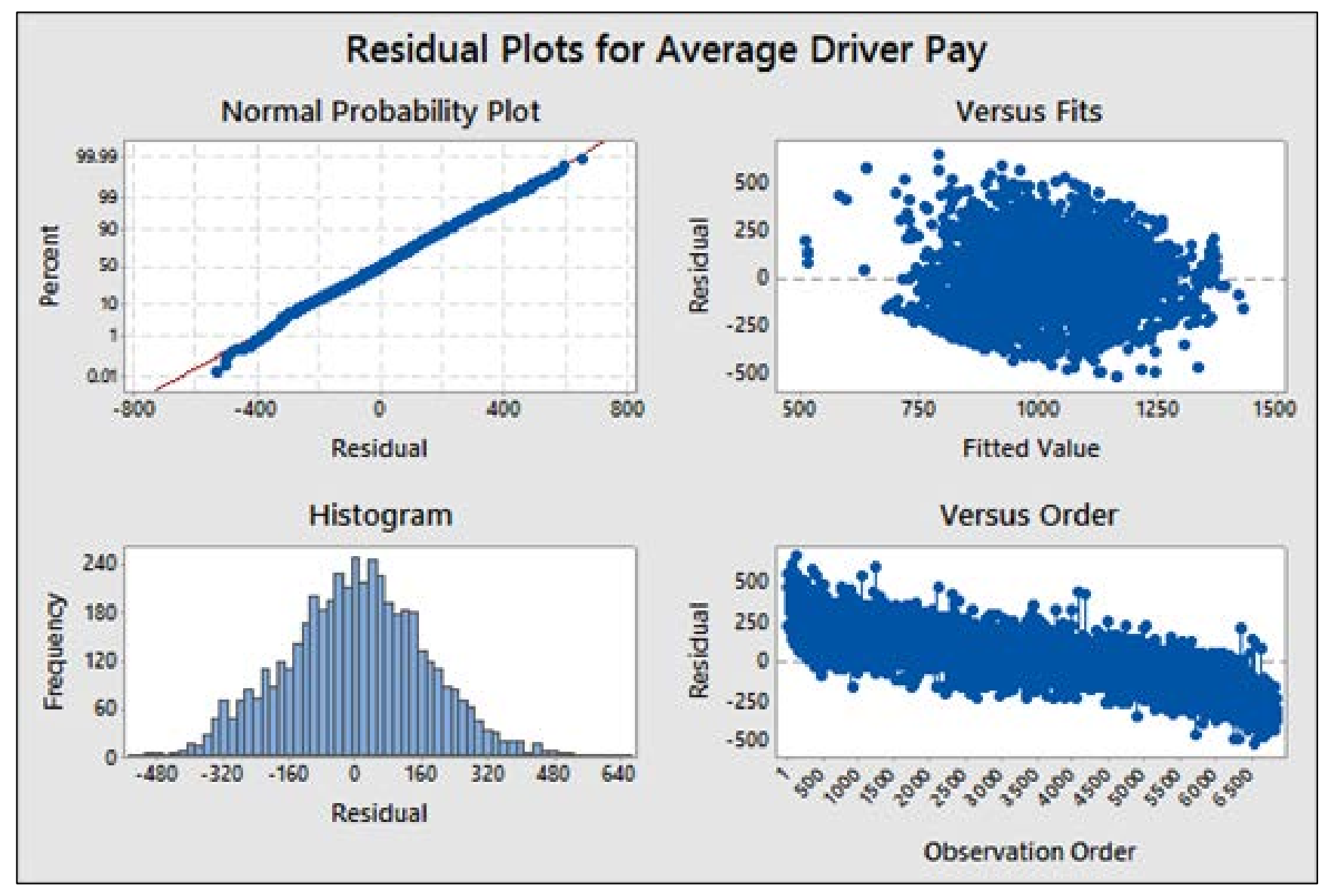
General Form

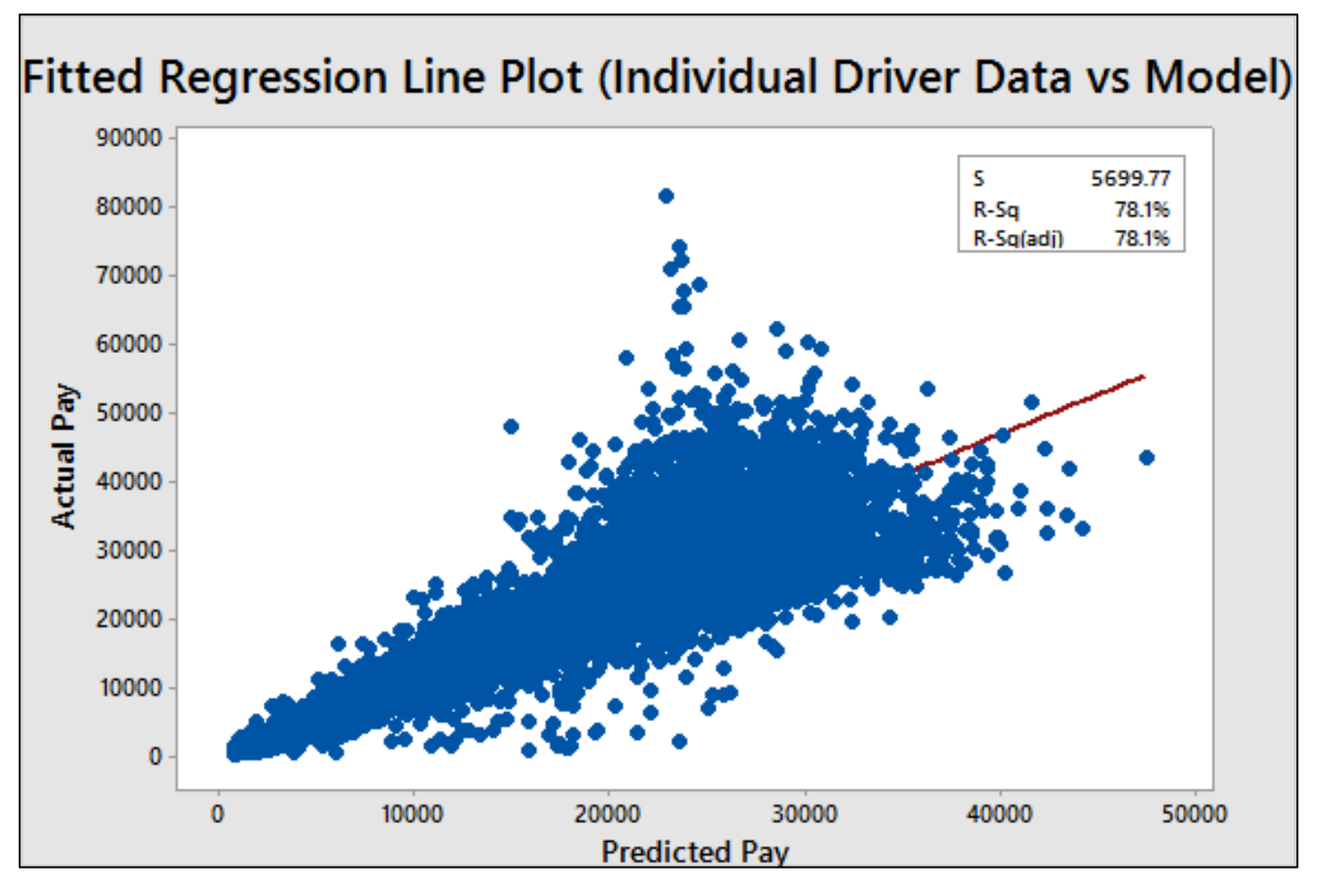
$$y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \cdots + \beta_{10} X_{10} + \varepsilon$$

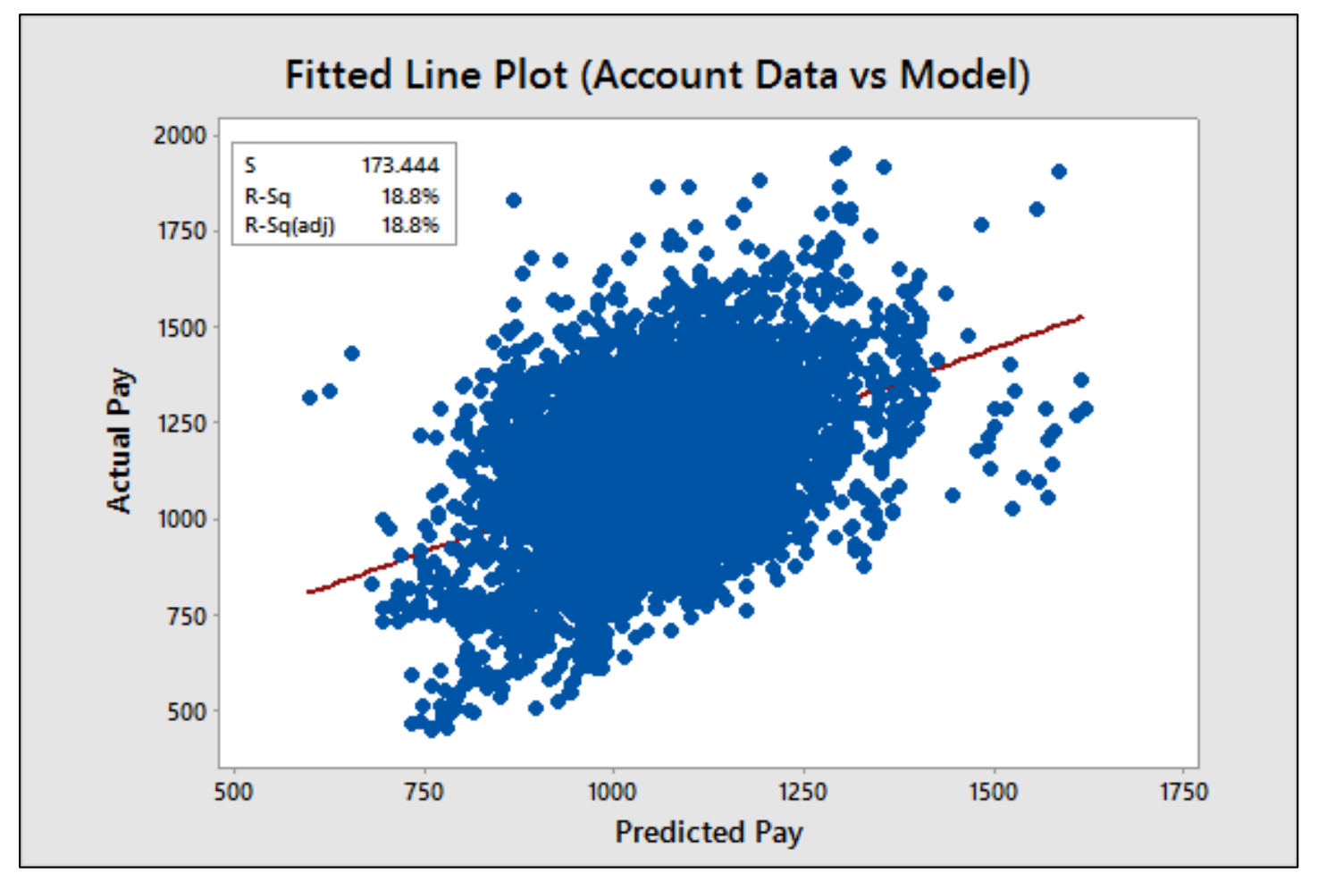
Verification

The model used provides the following residual plots suggesting normality among the predictions, as well as symmetrical distribution within the residuals vs. fits plot.

The multiple linear regression was tested against two sets of data, account level driver averages (testing set was in the same format) and against indiviudal driver data for the second set of 26 weeks.

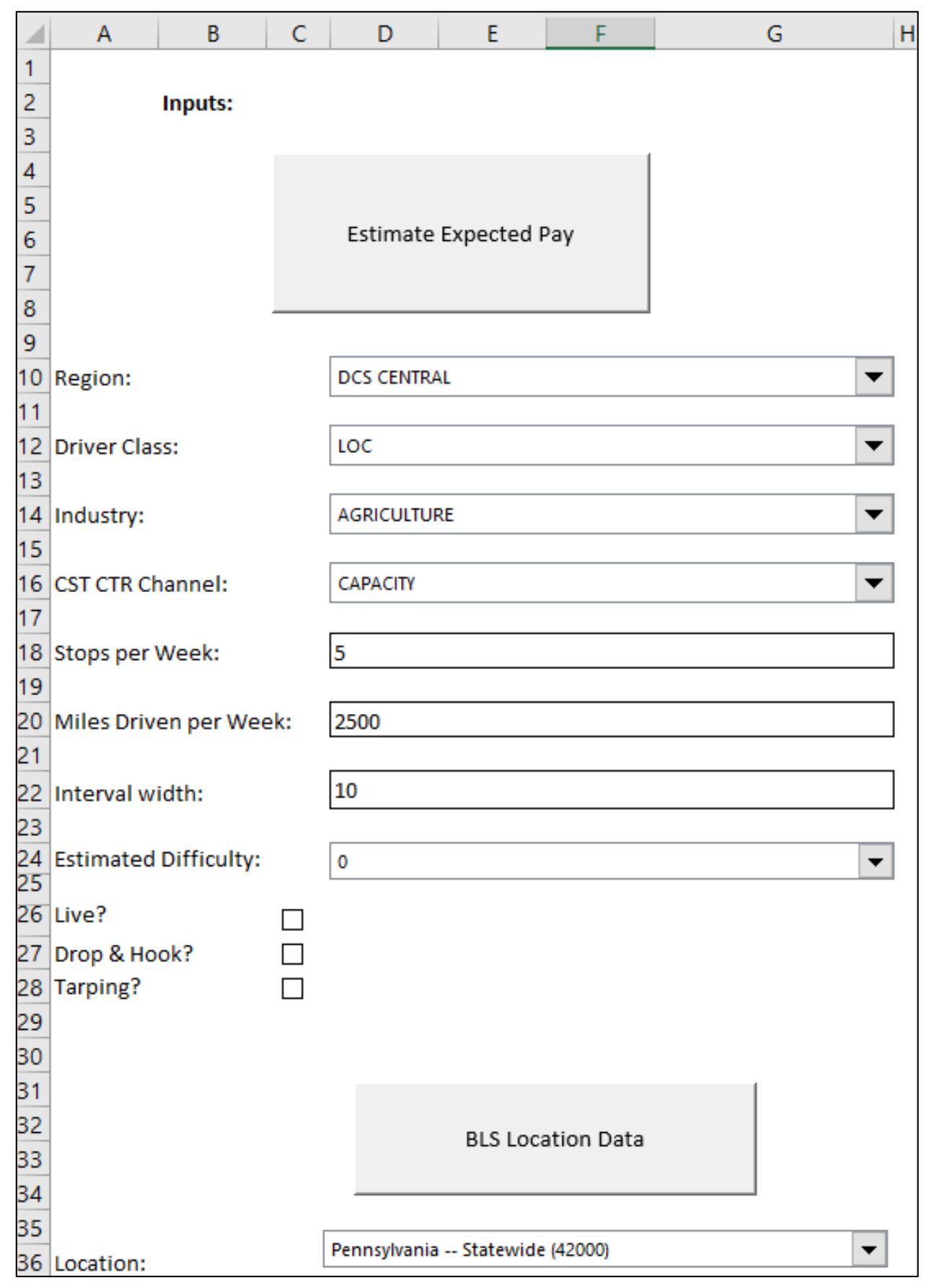

Residual Plots for Average Driver Pay


Fitted Regression Line Plot (Individual Driver Data vs Model)


Fitted Line Plot (Account Data vs Model)

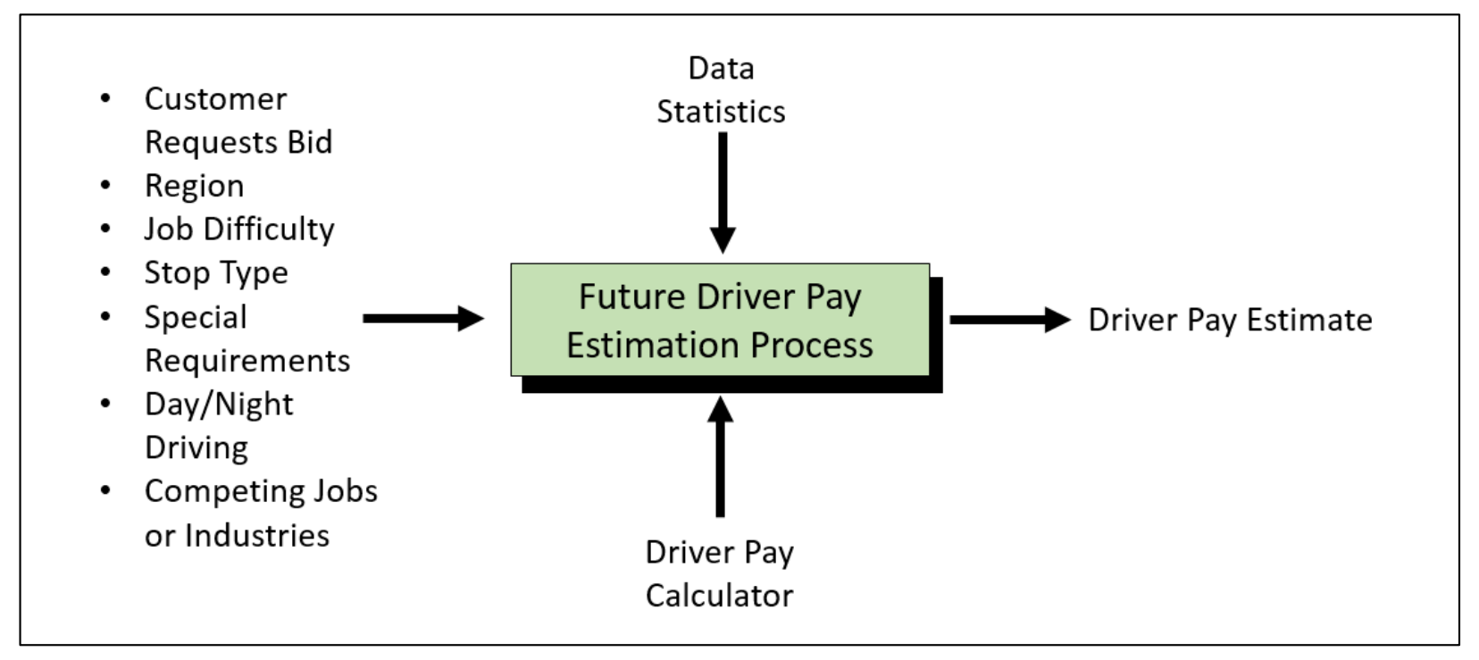
Decision Support Tool

The decision support tool has six functional areas. The first allows the user to give the tool inputs for the prediction. The inputs are auto-populated from the data set in the tool, so if the user loads a different dataset, the tool will not need to be manually updated. The inputs also remain on screen by design, so the user can reference the inputs when evaluating the prediction. The validation section of the tool is created by using J.B. Hunt's data to create a histogram displaying driver pay data for the selected region. For validation, our code retrieves information online from the Bureau of Labor Statistics, then processes the data to create a chart displaying the historical industry average for the selected area. The results section is created by computing the predicted pay from our data analysis. The data analysis produced an equation we use along with the user inputs. A range for the prediction is given, and the user can control the width of the range. We recognize there will be outside factors affecting the accuracy of the prediction the user wants to consider in the prediction, so we provided the adjustments section to allow the user to increase or decrease the prediction by a certain percentage. Finally, we display the accuracy of the ranged predictions based on the data within the tool. Each driver is used as inputs for our prediction, then each prediction is compared to the actual amount paid. The equation given below is used to calculate the accuracy displayed.



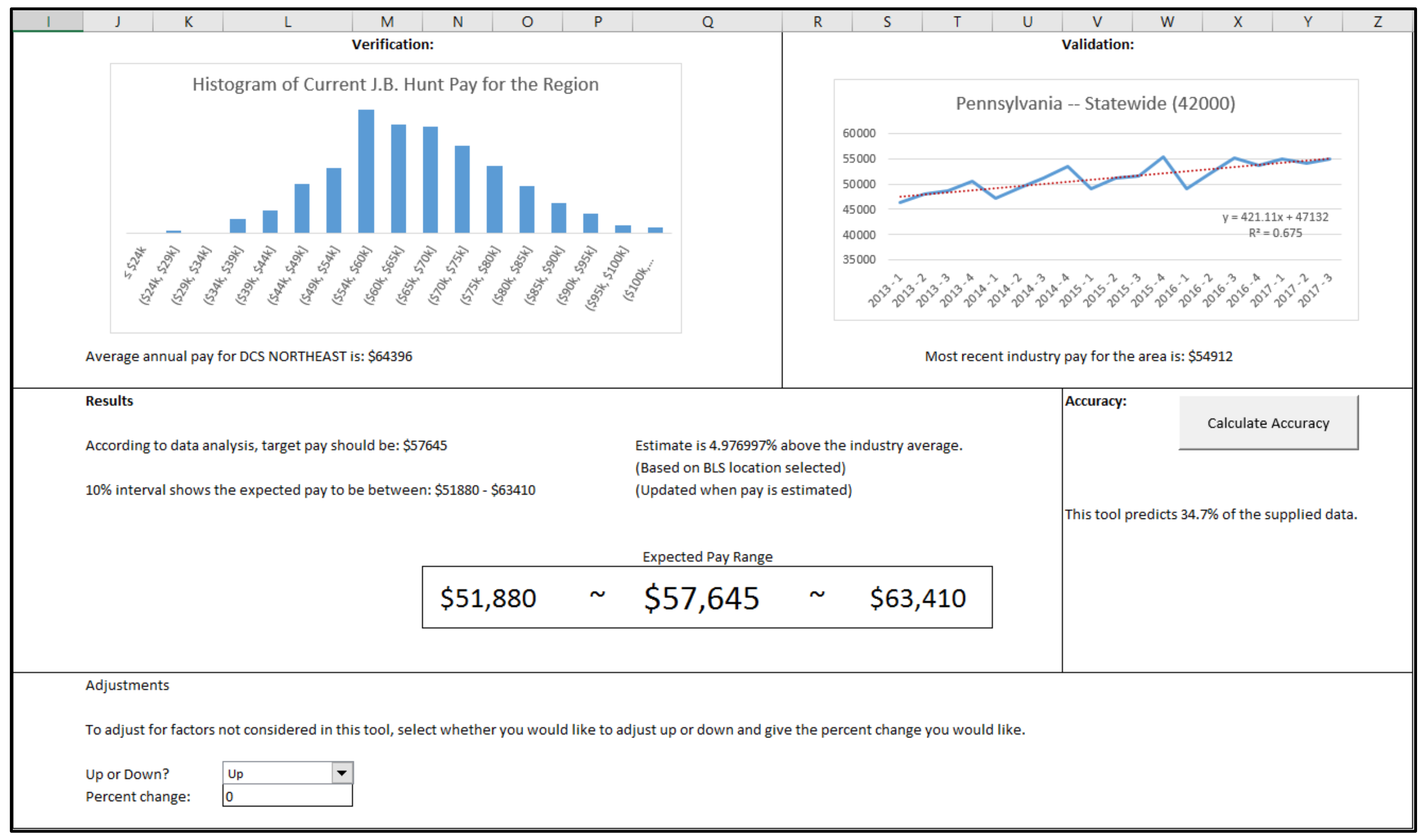
Inputs:

Region: DCS CENTRAL
Driver Class: LOC
Industry: AGRICULTURE
CST CTR Channel: CAPACITY
Stops per Week: 5
Miles Driven per Week: 2500
Interval width: 10
Estimated Difficulty: 0
Live? ☐
Drop & Hook? ☐
Tarping? ☐
Location: Pennsylvania -- Statewide (42000)



BUREAU OF LABOR STATISTICS
U.S. DEPARTMENT OF LABOR

$$Accuracy = \frac{\sum Accurate\ Predictions}{\sum All\ Predictions}$$



Verification:

Average annual pay for DCS NORTHEAST is: $64396

Validation:

Most recent industry pay for the area is: $54912

Results

According to data analysis, target pay should be: $57645

10% interval shows the expected pay to be between: $51880 - $63410

Estimate is 4.976997% above the industry average.
(Based on BLS location selected)
(Updated when pay is estimated)

Expected Pay Range

$51,880 ~ $57,645 ~ $63,410

Accuracy:

This tool predicts 34.7% of the supplied data.

Adjustments

To adjust for factors not considered in this tool, select whether you would like to adjust up or down and give the percent change you would like.

Up or Down? Up
Percent change: 0