Business Case

Anguie Garcia

1. Data modeling and integration diagrams.

CSV File kaggle	Simulated Source	Simulated Description	Reason for Source	Staging model snowflake
olist_marketing_qualified_leads_dataset.csv	HubSpot	Leads generated from marketing campaigns.	**HubSpot** is a widely-used CRM and marketing automation tool, ideal for tracking **Marketing Qualified Leads (MQLs)** generated through marketing efforts.	**stg_hubspot_leads**
olist_closed_deals_dataset.csv	Enrichment Data (Apollo)	Closed leads (conversion, CRM information, reps, etc.)	**Apollo** is a data enrichment platform that enhances lead data, making it suitable for tracking **closed deals**, customer profiles, and post-conversion interactions.	**stg_apollo_enrichment**
olist_orders_dataset.csv	Salesforce Data	Purchase orders related to sales.	**Salesforce** is a leading CRM for managing sales data, including **orders**. It's a central tool for sales teams to track customer purchases.	**stg_salesforce_orders**
olist_order_items_dataset.csv	Salesforce Data	Detailed product information in the orders.	**Salesforce** also tracks detailed order information, such as products sold, making it an accurate source for **order items**.	**stg_salesforce_order_items**
olist_order_payments_dataset.csv	Salesforce Data	Payments made by customers.	**Salesforce** records payment transactions, offering a clear view of customer payments and revenue, crucial for financial tracking and reporting.	**stg_salesforce_order_payments**

Table1. Simulation sources integration.

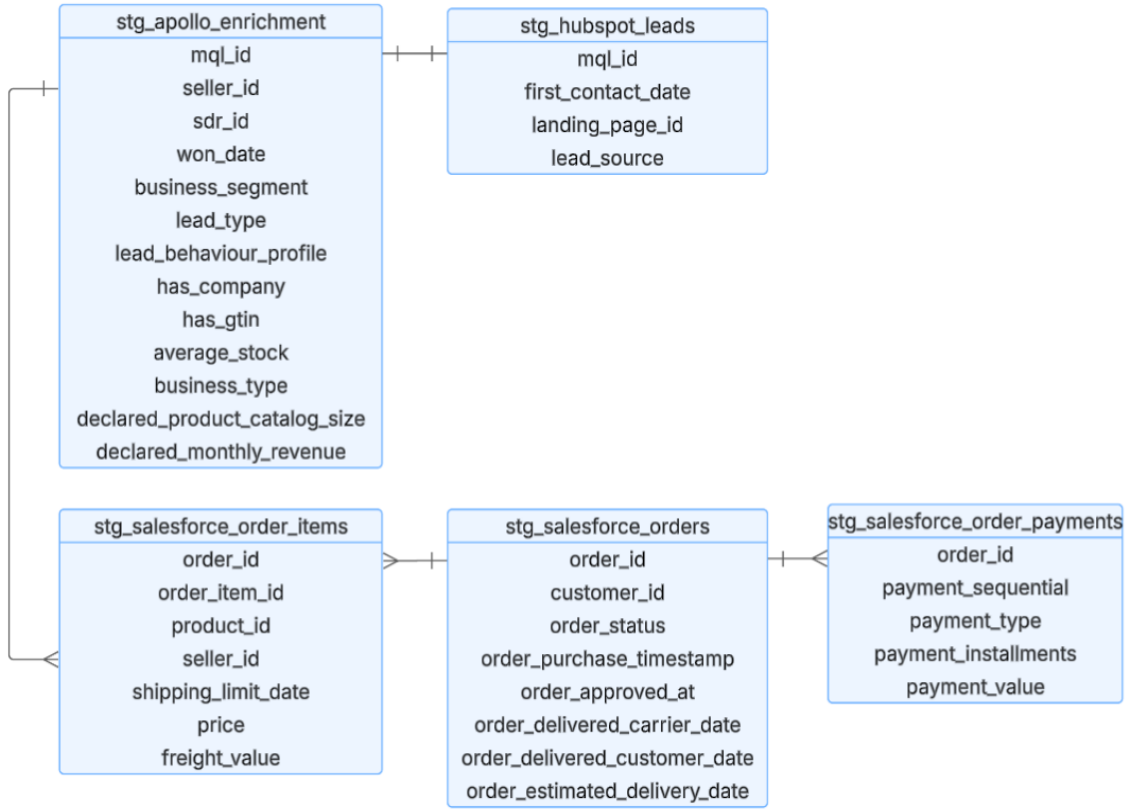
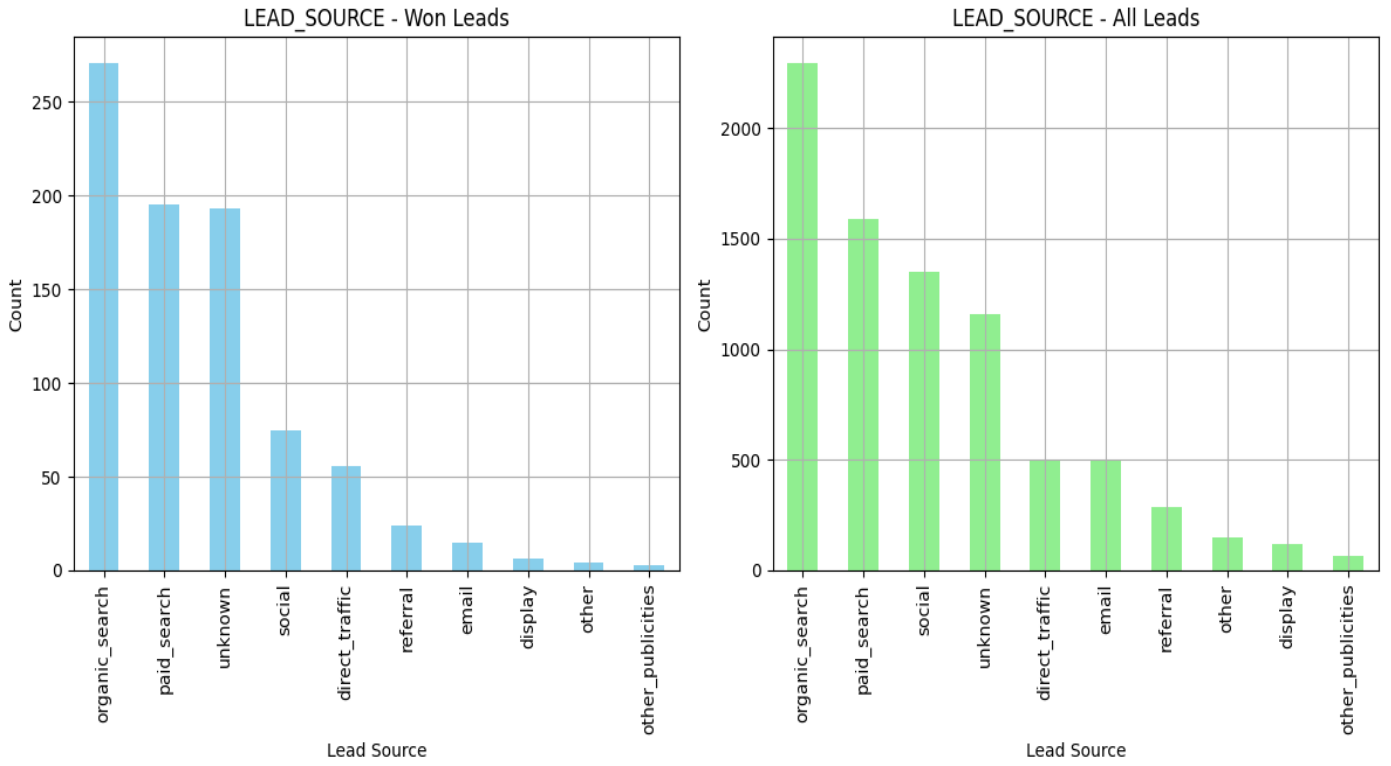


Fig1. Conceptual model.

2. Funnel & Attribution Analytics

Leads Funnel and conversion rates.

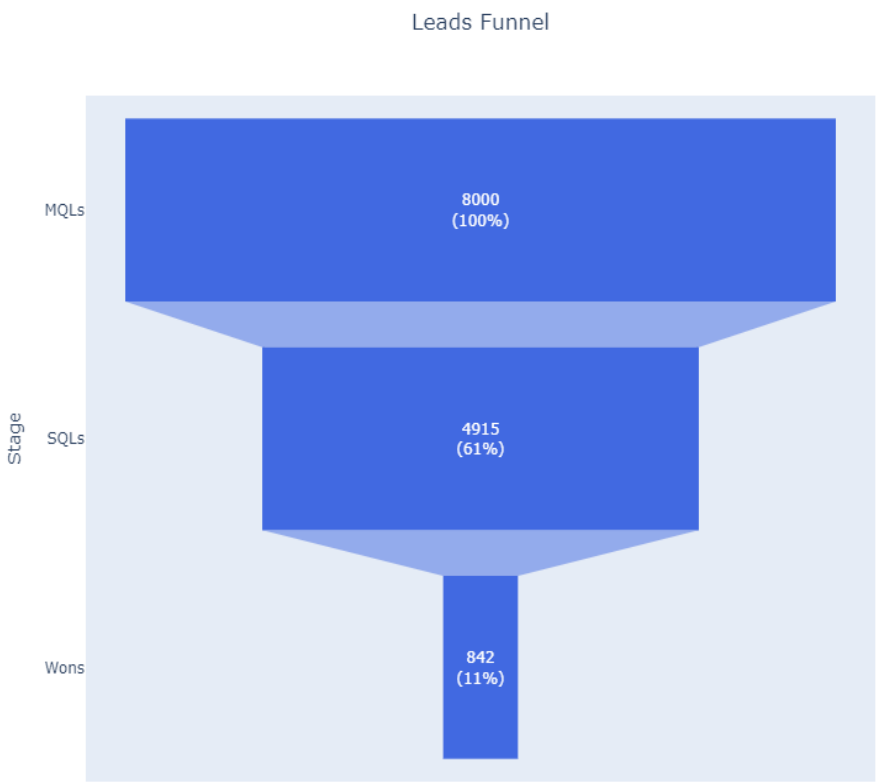


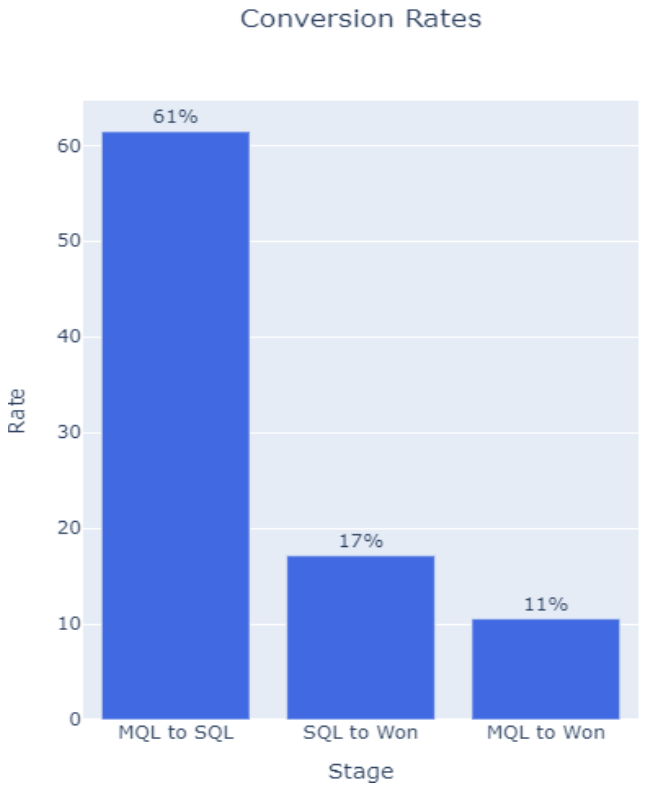
- The majority of visitors come from organic search (285), **indicating strong SEO performance and content relevance.**
- Paid search (195) is the second-largest source, showing effective advertising efforts, though it's important to monitor ROI to ensure cost efficiency.
- A notable 179 visitors have an unknown source, which warrants further investigation.
- Social media (75) and direct traffic (56) contribute modestly but play key roles in brand awareness and retention.
- Other channels—such as referrals, email, and display ads—show smaller impacts, yet they can be valuable in niche or targeted campaigns.

* Assumption for SQL: If landing_page_id ends with a number, it is ASSUMED TO BE CONTACTED (SQL).

2. Funnel & Attribution Analytics

Leads Funnel and conversion rates.





10.53% of MQLs convert into paying customers. Considering that over 61% of MQLs advance to SQLs, **this suggests that the biggest bottleneck lies in the sales phase.**

Marketing is generating a good number of MQLs, but not all of them are converting into customers. **Optimizing marketing campaigns** to improve the quality of the leads generated could enhance the conversion rate across the entire funnel.

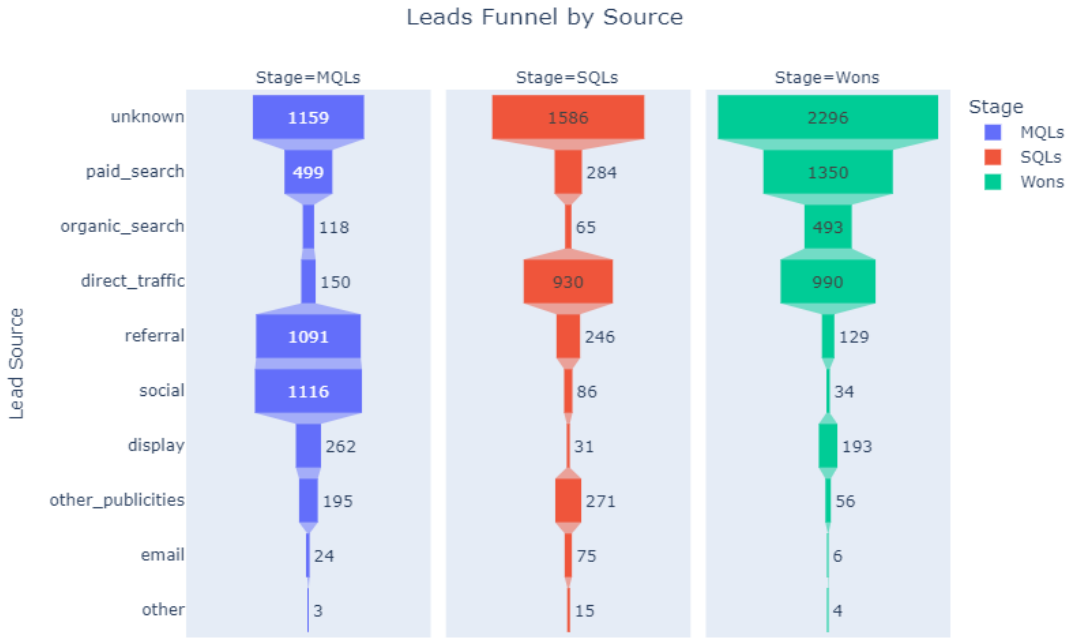
* Assumption for SQL: If landing_page_id ends with a number, it is ASSUMED TO BE CONTACTED (SQL).

2. Funnel & Attribution Analytics

Leads Funnel and conversion rates.

Source	MQL→SQL	SQL→WON	MQL→WON	🏆 Final Evaluation
unknown	✅ 80.24%	✅ 20.75%	✅ 16.65%	⭐ High priority – investigate and scale
paid_search	🔄 62.42%	🔄 19.70%	🔄 12.30%	⚠️ Medium-high – audit to optimize
organic_search	🔄 47.52%	✅ 24.84%	✅ 11.80%	⭐ High priority – invest more
direct_traffic	🔄 49.30%	✅ 22.76%	🔄 11.22%	➕ Medium – maintain and strengthen
referral	❌ 45.42%	🔄 18.60%	❌ 8.45%	➖ Low – reevaluate strategy
social	✅ 82.67%	❌ 6.72%	❌ 5.56%	⚠️ Low efficiency – improve lead quality
display	✅ 72.88%	❌ 6.98%	❌ 5.08%	🚫 Very low efficiency – reconsider investment
other_publicities	🔄 52.31%	❌ 8.82%	❌ 4.62%	❌ Very low – not profitable
email	🔄 53.14%	❌ 5.73%	❌ 3.04%	🚫 Very low efficiency – adjust or pause
other	❌ 20.67%	🔄 12.90%	❌ 2.67%	❌ Very low – not effective

Legend:
- ✅ Excellent performance
- 🔄 Medium or acceptable performance
- ❌ Low performance
- ⭐ Investment or improvement priority



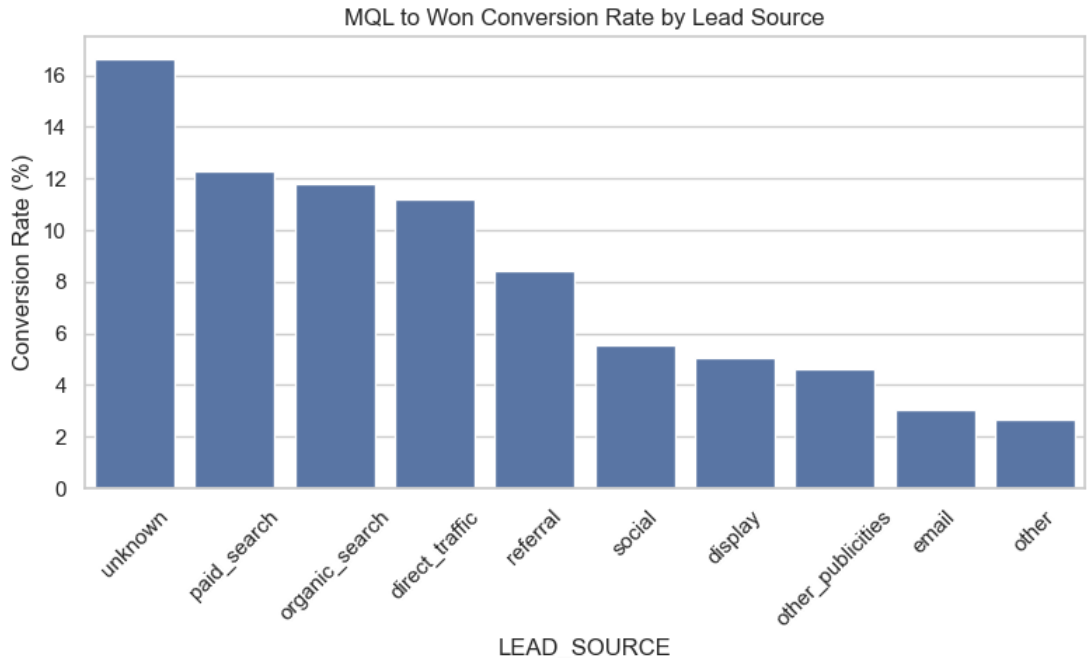
Recommendations

- Investigate what traffic is classified as "unknown" to correctly identify it and potentially scale it.
- Increase investment in SEO and organic content. Optimizing this channel can generate more returns without relying on direct investment.
- Audit paid campaigns. Adjust keywords, targeting, and copy to improve lead quality and conversion rates.
- Strengthen branding and enhance direct access (e.g., better CTAs in emails or social media).

* Assumption for SQL: If landing_page_id ends with a number, it is ASSUMED TO BE CONTACTED (SQL).

2. Funnel & Attribution Analytics

Leads Funnel and conversion rates.



Conclusions:

1. Channels with High Conversion Rates:

1. **Unknown** has the highest win rate (16.65%), but its unclear classification makes it difficult to act on. It may require cleanup to understand the source and optimize it effectively.
2. **Paid Search** is a strong performer with a win rate of 12.30%, delivering high volume and decent quality.
3. **Organic Search** and **Direct Traffic** also perform reasonably well, with win rates of 11.80% and 11.22%, respectively.

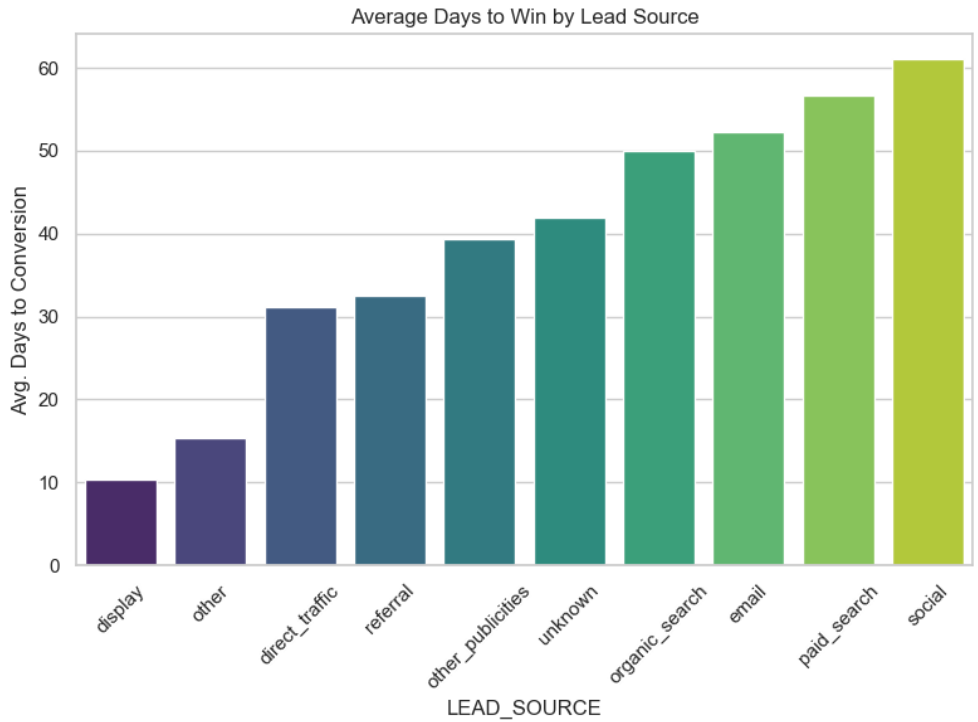
📌 **Conclusion**:

Although **Unknown** has the highest win rate, it is not actionable. The **top actionable channel** for scaling efforts is **Paid Search**, as it offers a good combination of **high volume** and **decent lead quality**.

* Assumption for SQL: If landing_page_id ends with a number, it is ASSUMED TO BE CONTACTED (SQL).

2. Funnel & Attribution Analytics

Leads Funnel and conversion rates.



Conclusions:

2. Channels with Fast Closing Times:

1. **Direct Traffic** (31.1 days) and **Referral** (32.5 days) have the **fastest closing times** among the channels, despite their lower total conversion rates.
2. **Unknown** (41.9 days) and **Organic Search** (50.0 days) take a bit longer to close deals.
3. **Paid Search** has the slowest closing time at 56.6 days.

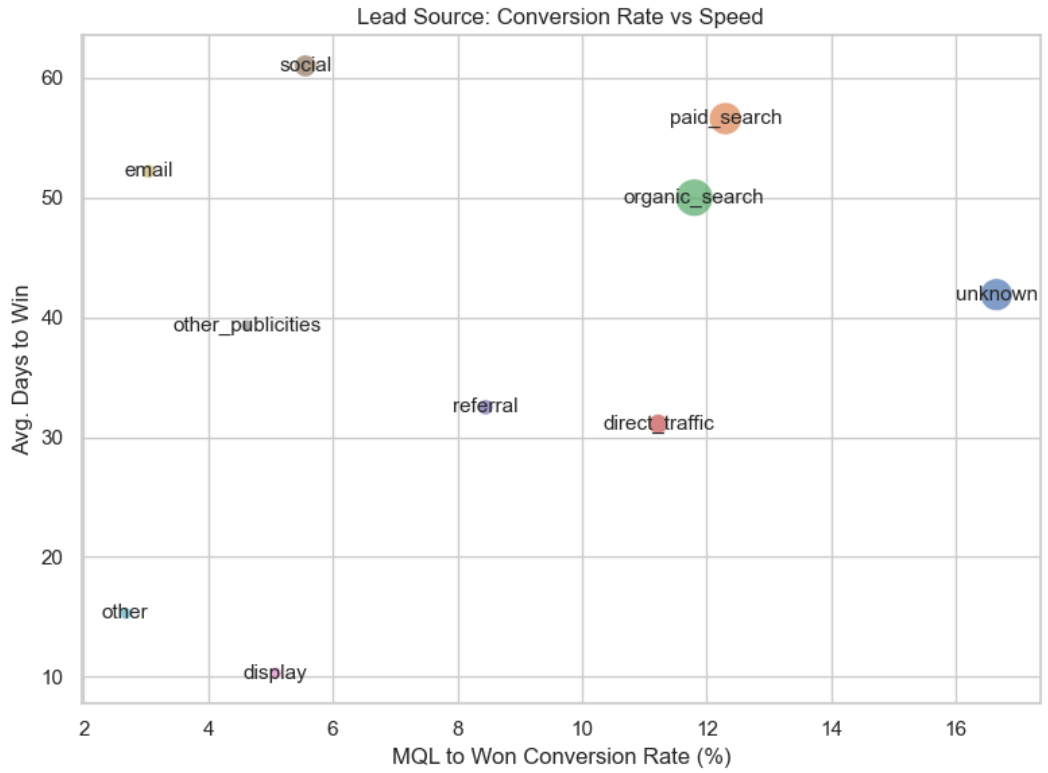
📌 **Conclusion**:

Direct Traffic and **Referral** close deals the fastest, making them valuable for quick wins, even though they convert fewer leads in total compared to other channels.

* Assumption for SQL: If landing_page_id ends with a number, it is ASSUMED TO BE CONTACTED (SQL).

2. Funnel & Attribution Analytics – First Attribution

Attribution analytics – first attribution



🎯 **Key Conclusions**

Best Sources for Conversion:

- **Best channels by value (conversion rate):**
  unknown, paid_search, organic_search, and **direct_traffic** are the most effective channels in terms of conversion.
- **Fastest channels:**
  display, direct_traffic, referral, and **other** convert in less than 35 days on average, which can be useful for accelerating revenue.
- **Least effective channels:**
  email, social, other_publicities, and **other** have low conversion rates and/or high conversion times, meaning they require intervention or de-prioritization.

Action	Channel
✅ **Prioritize by effectiveness**	unknown, paid_search, organic_search
⚡ **Leverage by speed**	direct_traffic, display, referral
🧪 **Review strategy**	social, other_publicities, email
🛑 **Consider budget cuts**	other, email

* Assumption for SQL: If landing_page_id ends with a number, it is ASSUMED TO BE CONTACTED (SQL).

2. Funnel & Attribution Analytics – Logistic Regression

Attribution analytics – linear regression

Lead Source	Model Coefficient	Impact on Conversion	Evaluation
unknown	**+0.663**	High Positive	✅ Strong conversion channel
paid_search	**+0.548**	Moderate Positive	👍 Effective for continuous investment
organic_search	**+0.365**	Moderate Positive	👍 Good organic performance
referral	**+0.109**	Low Positive	🟡 Limited potential
other_publicities	**−0.399**	Moderate Negative	⚠️ Low effectiveness
social	**−0.449**	Moderate Negative	⚠️ Low return
other	**−0.533**	Clear Negative	🚫 Not recommended
email	**−0.882**	Strong Negative	🚫 Avoid or review strategy

🎯 **Key Conclusions**

Best Sources for Conversion:

•**Unknown**, **Paid Search**, and **Organic Search** are the most positively influential channels.

•These should be maintained or receive increased investment and monitoring.

Sources to Reevaluate:

•**Email**, **Other**, **Social**, and **Other Publicities** have negative coefficients.

•This could indicate low lead quality or misalignment with the message. These channels require optimization or budget reduction.

Referrals have a low but positive impact: you could experiment with improvements in this channel (referral programs, rewards, etc.).

Action	Channel
📈 **Increase Investment**	unknown, paid_search, organic_search
🧪 **Optimize Strategy**	referral, social, other_publicities
🛑 **Reduce Budget or Review Approach**	email, other

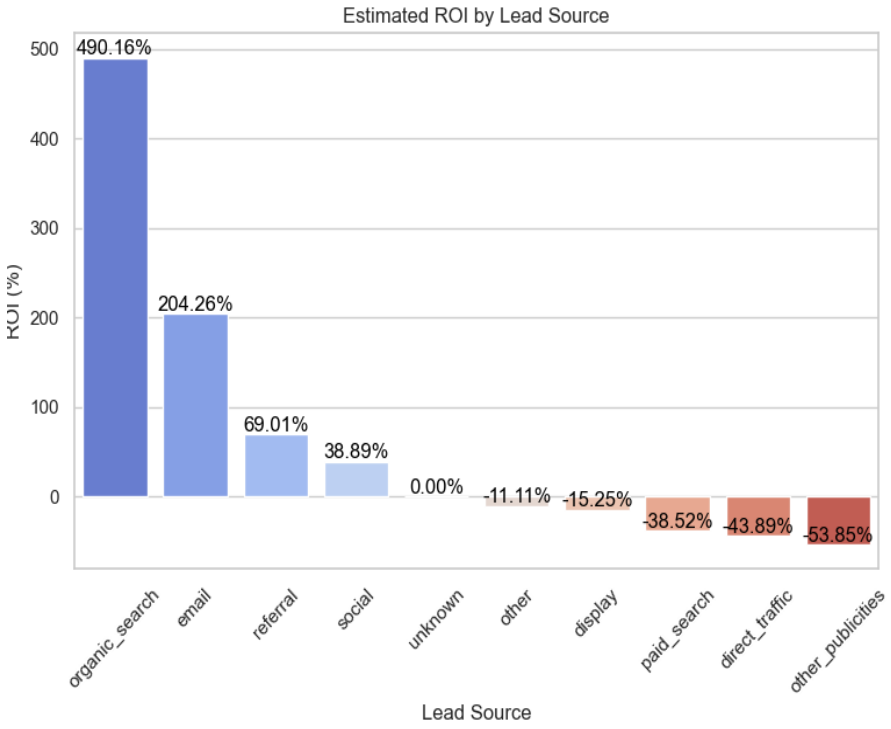
* Assumption for SQL: If landing_page_id ends with a number, it is ASSUMED TO BE CONTACTED (SQL).

2. Funnel & Attribution Analytics - ROI

ROI

📌 **Hypothetical Assumptions:**

- Average revenue per **Won lead** = $1,000
- Cost per lead by source:
  - **Paid Search** = $100
  - **Organic Search** = $10
  - **Referral** = $25
  - **Social** = $20
  - **Email** = $5
  - **Unknown** = $0 (Assume organic/untracked)
  - **Display** = $30
  - **other_publicities** = $50
  - **Other** = $15
  - **direct_traffic** = = $100



🎯 **Strategic Analysis**

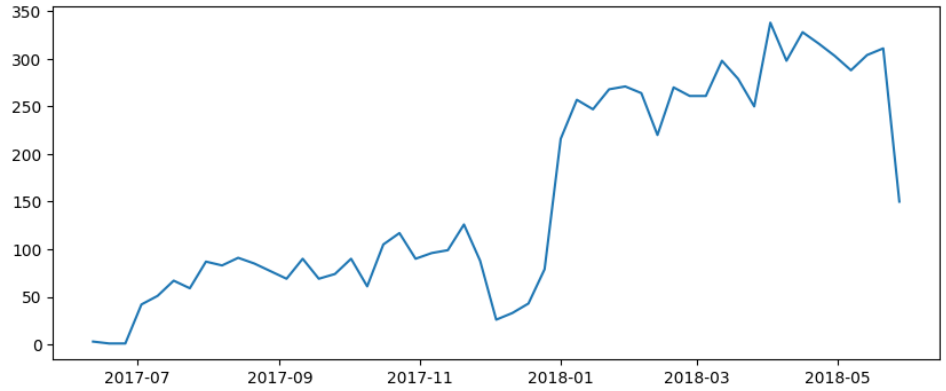
- Organic Search and Email not only convert well but also generate significantly more revenue than they cost.
- Channels like Paid Search, despite having good conversion rates, are not profitable, indicating that the cost per acquisition is too high.
- Direct Traffic has a good conversion time but destroys economic value, which could suggest that it is either not effectively influenced by campaigns or not properly attributed.
- Social and Referral could be maintained with adjustments if campaigns or costs are optimized.

Action	Channel
✅ **Increase Investment**	Organic Search, Email
🧪 **Maintain with Adjustments**	Referral, Social
⚠️ **Review Strategy and Costs**	Paid Search, Direct Traffic, Display
🛑 **Consider Budget Cut**	Other Publicities, Other

* Assumption for SQL: If landing_page_id ends with a number, it is ASSUMED TO BE CONTACTED (SQL).

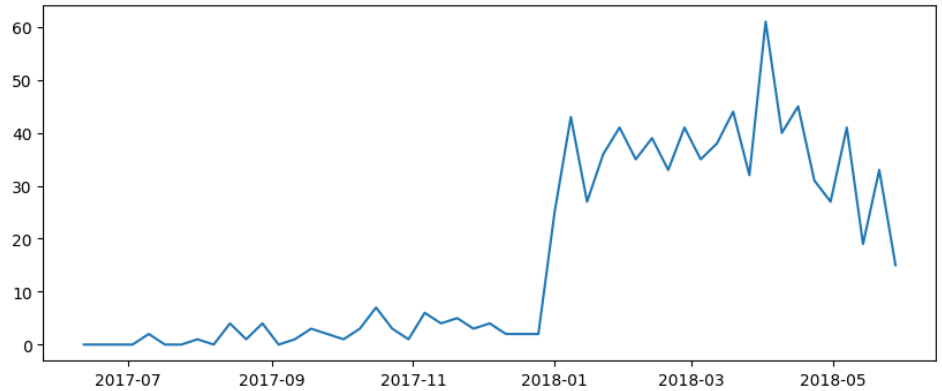
3. Forecasting & Planning

Leads



- From June to December 2017, the number of leads increased in a moderate but steady manner.
- Starting from January 2018, there is a sharp jump in the number of leads, rising from around 100 to more than 250 per week.
- The peak is reached in April 2018, surpassing 320 leads per week.
- Abrupt decline in the last observed week (late May 2018), a significant drop is noticeable.

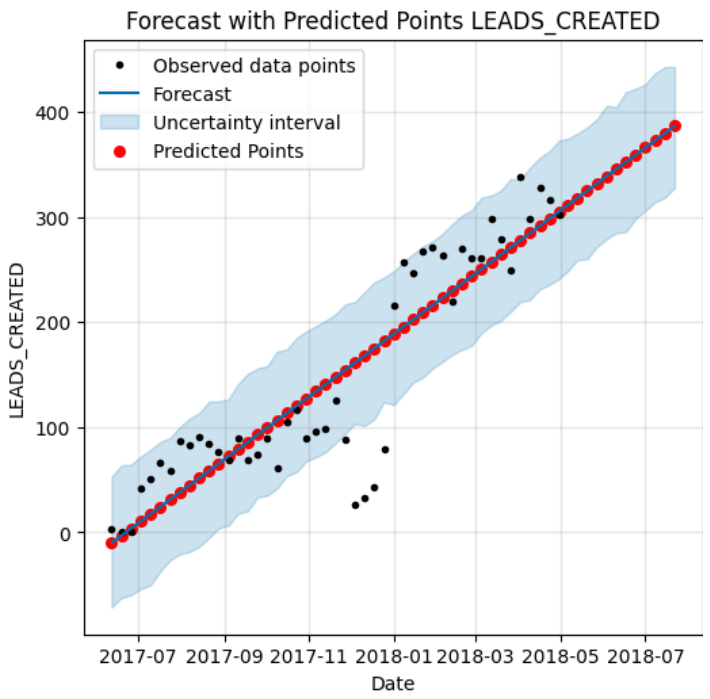
Conversions

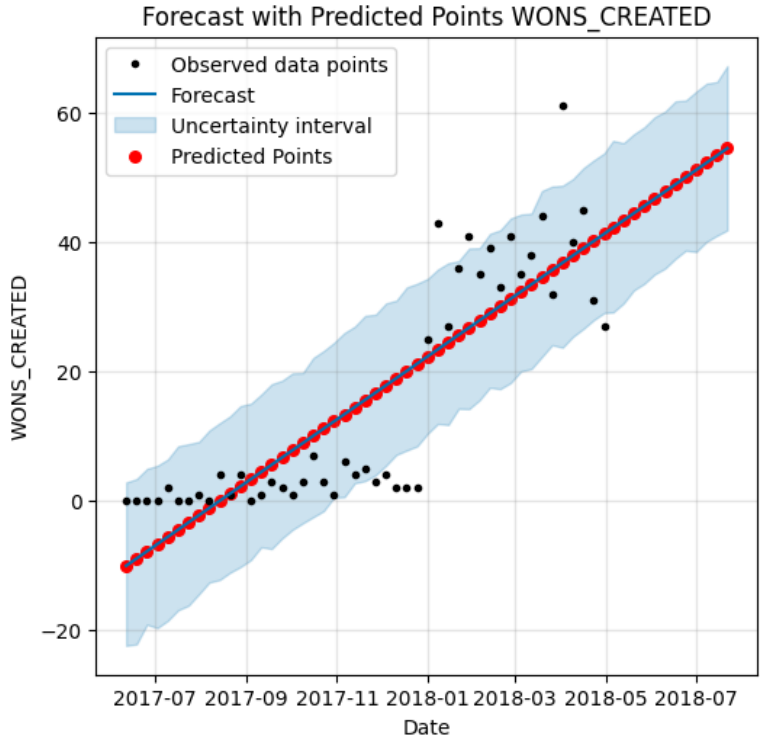


- The chart reveals a sharp spike in the data points between November 2017 and January 2018, peaking at over 60 in early 2018, followed by a decline and fluctuating behavior through mid-2018.
- The pattern suggests a potential event or change around late 2017 that triggered the dramatic increase, with a subsequent stabilization after the peak.

* Assumption for SQL: If landing_page_id ends with a number, it is ASSUMED TO BE CONTACTED (SQL).

3. Forecasting & Planning - Prophet

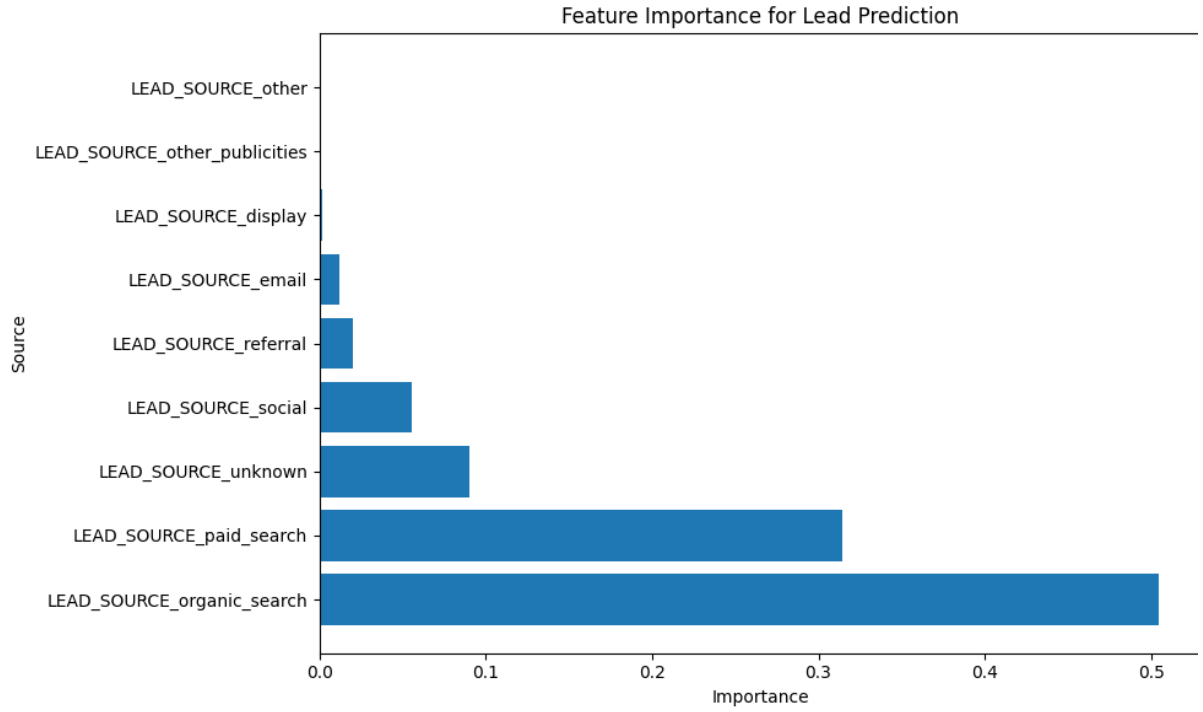




- Observed Data Points (black dots): These are the actual values of LEADS_CREATED or WONS over time. They show some fluctuation but generally follow an upward trend.
- Forecast (blue line): This represents the model's prediction for future values. It continues the upward trend, suggesting increasing leads over time.
- Uncertainty Interval (shaded blue area): This shows the model's confidence in its predictions. The wider the interval, the more uncertainty the model has. The forecast becomes slightly more uncertain over time, as expected.
- Predicted Points (red dots): These are the specific predicted values over time, which lie on the blue forecast line.

* Assumption for SQL: If landing_page_id ends with a number, it is ASSUMED TO BE CONTACTED (SQL).

3. Forecasting & Planning - regression



The feature importance values show how much each lead source contributes to the model's ability to predict the total number of leads per week.

- Organic search is by far the most important source, accounting for over 50% of the model's predictive power.
- Paid search follows with around 31%, indicating it also plays a major role.
- Unknown sources (9%) and social media (5.5%) have moderate influence, while referrals, email, and display ads contribute very little.
- The sources other publicities and other have negligible impact on the model, suggesting they do not significantly influence weekly lead totals in the current dataset.

* Assumption for SQL: If landing_page_id ends with a number, it is ASSUMED TO BE CONTACTED (SQL).

4. Insights and Recommendations

Insights

- **61% of MQLs advance to SQLs, but only 10.53% convert to customers**

- **Channels with the Highest Impact on Conversion and Value.**
  Organic Search and *Email* are the most profitable channels. *Paid Search* converts well but loses money, and *Unknown* needs to be investigated due to its high unassigned performance.

Actionable Recommendations

- **Increase investment in:** *Organic Search* (Enhance SEO efforts and publish high-quality, relevant content consistently. This is the strongest lead source and deserves further optimization and investment), *Email* (Although it currently has low importance, improving segmentation, personalization, and calls to action can help unlock its potential)
- **Optimize campaigns in:** *Paid Search* (Fine-tune keyword strategies, improve ad quality, and refine audience targeting to increase efficiency and ROI), *social* (Reassess budget allocation and optimize targeting strategies. Focus on high-performing platforms and reduce unnecessary spend.)
- **Reduce or eliminate budget for:** *Display, Other Publicities, Other, Direct Traffic*. These sources show minimal or no contribution to lead prediction. Budgets here should be cut or redirected toward more impactful channels.
- **Conversion Funnel Optimization.** Strengthen the post-SQL sales process with better tools, follow-up, or training.
- **Data Cleanliness and Attribution.** Investigate what traffic is classified as "Unknown" for optimization.

Next steps

Reorganize Snowflake Connection Parameters:

- Reorganize the connection parameters in Snowflake to ensure they are properly structured with keys, rather than hardcoded in the 3 notebooks.

Use Incremental Models for Record Updates:

- Implement incremental models to automatically update records whenever new leads and conversions are added. This ensures that the data remains up-to-date without requiring full refreshes.

Integrate Order and Payment Information:

- Merge the data from the orders and order_payments tables to demonstrate how the generated leads (won leads) contribute to real revenue.
- Collaborate with the sales team to review the revenue data by salesperson, ensuring accuracy and clarity.

Define Data Sources in sources.yml:

- Update the sources.yml file to define the data sources for HubSpot and Apollo

Validate Model Performance:

- It is possible for RMSE and MAPE to increase.
- Use models like SARIMA
- Contribution of Each Lead Source per Week for conversions.